BOSSA: Learning Music Style Through Cross-Modal Bootstrapping

Jingwei Zhao[1,3] **Ziyu Wang**[4,5] **Gus Xia**[5] **Ye Wang**[2,1,3]

[1]Institute of Data Science, NUS [2]School of Computing, NUS
[3]Integrative Sciences and Engineering Programme, NUS Graduate School
[4]Courant Institute of Mathematical Sciences, NYU [5]Music X Lab, MBZUAI

Abstract

What is music style? Though often described using text labels such as "swing," "classical," or "emotional," the real style remains implicit and hidden in concrete music examples. In this paper, we introduce a cross-modal framework that learns implicit music styles from raw audio and applies the styles to symbolic music generation. Inspired by BLIP-2, our model leverages a Querying Transformer (Q-Former) to extract style representations from a large, pre-trained audio language model (LM), and further applies them to condition a symbolic LM for generating piano arrangements. We adopt a two-stage training strategy: contrastive learning to align style representations with symbolic expression, followed by generative modeling to perform music arrangement. We name our model as BOSSA (i.e., BOotStrapping audio-to-Symbolic Arrangement). It generates piano performances jointly conditioned on a lead sheet (content) and a reference audio example (style), enabling controllable and stylistically faithful arrangement.[1]

1 Introduction

Automatic music generation is often controlled by *explicit* content such as melody, chords, and text labels [2, 16, 25, 30], but music concepts can be more nuanced than we often realize. When musicians learn a style, instead of relying on abstract definitions like "romantic" or "jazz" alone, they absorb patterns from music examples that share common stylistic traits. The commonality across these examples forms a style, an *implicit* one that cannot be fully described with words or labels but only understood through the music itself. This paper explores how such implicit style can be internalized from music examples and used to control music generation in a deep learning framework.

Large-scale music language models (music LMs) have shown strong capabilities in learning explicit music content, as demonstrated by probing studies [3, 17, 18, 23, 27] and adapter-based designs [13, 14, 29, 32]. Yet, control over implicit style remains limited. For example, when using audio to guide symbolic music generation, existing models can extract melody and chords [7, 26], but capturing stylistic traits like comping patterns or voicing preferences remains a greater challenge. This requires disentangling style from music content, which current music LM-based studies have yet to explore.

In this paper, we explore learning implicit music style in a cross-modal setting for symbolic piano arrangement. Our goal is to generate an arrangement conditioned on two inputs: an audio example (providing style) and a lead sheet (melody and chords as content). To achieve this, we connect pre-trained music LMs in the audio and symbolic domains using a Querying Transformer (Q-Former), a lightweight Transformer originally designed for vision-language alignment [12]. As shown in Figure 1, we extend its role to capture implicit music style, extracting a *style representation* from

[1]Demo page: `https://zhaojw1998.github.io/bossa/`

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI for Music.

the hidden states of the audio LM. The symbolic LM then conditions on this representation, along with the *content* of the lead sheet, to generate an arrangement. The Q-Former enables style transfer between two large unimodal LMs without re-training them—a process we refer to as *bootstrapping*.

In our design, we treat the Q-Former as a bottleneck to transfer only style-related information and adopt a two-stage training strategy. The first stage employs contrastive learning, training the Q-Former to extract auditory representations that are musically relevant, expressible in symbolic piano composition, and independent of explicit music content. The second stage focuses on generative modeling, where the Q-Former's output conditions the symbolic LM to arrange the desired piano performance. We show that the complete system generates more stylistically accurate cover songs compared to existing audio-to-symbolic arrangement methods, besides achieving piano style transfer by specifying audio examples.

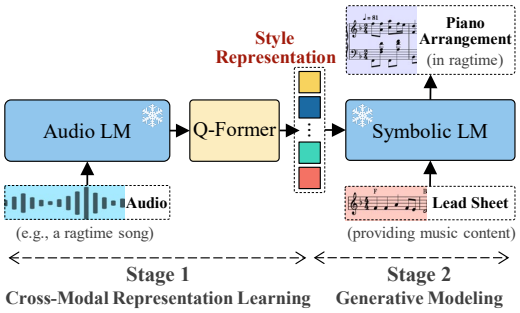


Figure 1: A cross-modal bootstrapping method.

In summary, the contributions of this paper are threefold:

1. We use the Q-Former to **align audio and symbolic modalities through implicit music style**, extending its role beyond content alignment in vision-language tasks.
2. We present **a new methodology to disentangle music style from large, pre-trained LMs**, offering a more scalable alternative to traditional latent-variable disentanglement methods.
3. Our model achieves **style-aware audio-to-symbolic piano cover arrangement**. Experiments demonstrate that it outperforms existing audio-to-symbolic models including both disentanglement-based methods and standard LM approaches.

2 Method

To bridge the modality gap from audio to symbolic music, we adopt the Q-Former [12] under a two-stage training strategy. Stage 1 focuses on audio-symbolic representation learning with a frozen audio LM, while Stage 2 addresses audio-to-symbolic arrangement with a symbolic LM. We introduce the Q-Former architecture and Stage 1 in Section 2.1, followed by Stage 2 in Sections 2.2.

2.1 Stage 1: Audio-Symbolic Representation Learning with Q-Former

The Q-Former is a Transformer encoder with two parallel, modality-specific streams that share the self-attention layers. As shown in Figure 2, it accepts both audio and symbolic inputs and learns a cross-modal music style representation. The left stream interacts with the audio LM to extract auditory music features. The right stream encodes symbolic music tokens. A set of querying embeddings (*queries*), randomly initialized, is fed to the left stream and serves as the bridge between the two modalities. At test time, the left stream is retained to extract cross-modal style directly from audio.

We integrate the Q-Former into MusicGen [4], one of the leading audio LMs available today. The queries interact with the audio hidden states via cross-attention, while remaining connected to the symbolic stream through the shared self-attention layers. To regulate cross-modal interactions, we employ tailored self-attention masks corresponding to specific training objectives, as detailed below.

Audio-Symbolic Contrastive Learning: We aim to enforce a higher audio-symbolic similarity for positive pairs compared to negative ones. Let $\mathbf{Z} \in \mathbb{R}^{K \times 768}$ be a sequence of K query outputs from the audio stream of Q-Former, and $t \in \mathbb{R}^{1 \times 768}$ be the output embedding of the start token (<s>) from the symbolic stream. We define the audio-symbolic similarity as $\max_k(\cos(\mathbf{Z}_k, t))$ for $k = 1, 2, \cdots, K$, where $\cos(\cdot, \cdot)$ is the cosine similarity. This contrastive loss pulls closer aligned audio and symbolic clips in the representation space while pushing apart unrelated pairs. To prevent information leakage, we employ a *unimodal self-attention mask*, ensuring queries and symbolic tokens do not attend to each other. For the detailed mask design, we refer readers to BLIP-2 [12].

Audio-Symbolic Matching: We also formulate a binary classification objective, where the model predicts whether a given audio-symbolic pair corresponds to each other. On top of the contrastive

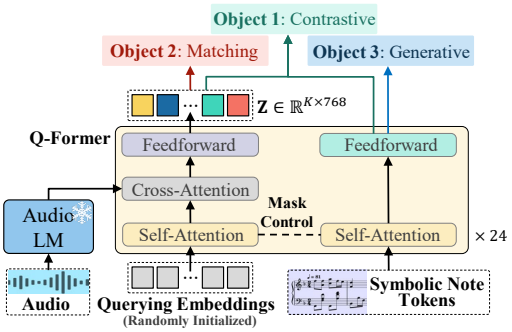


Figure 2: The Q-Former architecture.

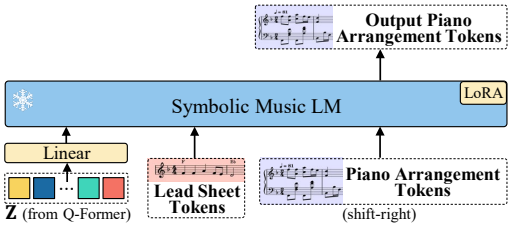


Figure 3: Audio-to-symbolic generative modeling prompted by the Q-Former outputs.

loss, the matching loss captures a finer cross-modal correspondence. We apply *no masking*, allowing the queries to attend across modalities. Each query output $\mathbf{Z}_k$ is fed into a binary linear classifier to produce a logit, and the logits from all queries are averaged to compute the final matching score.

Audio-Grounded Symbolic Generation: We further supervise the symbolic stream to auto-regressively predict piano arrangement tokens given the audio. We implement a *cross-model causal self-attention mask*, allowing the symbolic stream to see the queries but not vice versa. This generative loss enforces alignment between the query-extracted auditory style and its symbolic realization.

2.2 Stage 2: Audio-to-Symbolic Generative Modeling

In the generative modeling stage, we take advantage of the generative capability of MuseCoco [16], a symbolic music LM. As shown in Figure 3, MuseCoco is used to reconstruct a piano arrangement based on two concatenated conditional inputs: 1) the query output embeddings $\mathbf{Z}$ from the Q-Former, and 2) a lead sheet. The Q-Former is pre-trained at Stage 1 to extract cross-modal music style from the audio, thus providing style guidance. The lead sheet defines the melody and chord as the content.

To enable compatibility with MuseCoco, we project $\mathbf{Z}$ into the same embedding dimension as MuseCoco's token embeddings using a linear layer. Symbolic note tokens are represented in the REMI [10] format. Since MuseCoco does not natively support lead sheet conditioning and the inclusion of the lead sheet alters its input format, we incorporate a LoRA adapter [9] into each self-attention layer. This allows the model to reweight attention and accommodate the added conditioning inputs. Notably, MuseCoco itself remains frozen throughout this process.

3 Experiments

Our model takes a lead sheet and an audio reference as input. When the lead sheet and audio are paired, the task is *piano cover generation*; when not paired, the task is *style transfer*. Appendix A shows qualitative demos for the latter case. Given that prior work has primarily focused on piano cover generation, we benchmark our model on this task in this section. Our baseline models include *Picogen2* (PCG2) [21] and *Audio2MIDI* (A2M) [26]. To ensure a fair comparison, we use Sheetsage [6, 7] to transcribe lead sheets from the audio, making audio the sole input for all methods. We also compare against an ablation variant model, tagged *Ours w/o Pre-Training* (w/o PT), where the Q-Former is trained directly in Stage 2 without representation learning in Stage 1. We consider two evaluation datasets: an *in-distribution* set containing 46 samples of the POP909 test split [24], and an *out-of-distribution* set containing 100 tracks randomly drawn from the Ballroom [8, 11] and the GTZAN [19, 22] datasets. POP909 is part of the training data, which comprises pop music only. Ballroom/GTZAN contains more diverse genres and instrumentation that are unseen during training, which assesses the models' capability to accommodate *more general styles beyond pop music*. See Appendix B for details on the datasets, and Appendix C for model configuration and training details.

3.1 Objective Evaluation

We first evaluate our model's performance in terms of *audio-to-symbolic coherence*, specifically assessing how well the generated piano covers preserve the structures from the original audio. To do

	POP909		Ballroom/GTZAN	
	Acc@5 (%) ↑	Rank ↓	Acc@5 (%) ↑	Rank ↓
Ours	**27.1** ± 1.2	**16.3** ± 0.3	21.4 ± 0.6	**30.2** ± 0.3
PCG2	23.6 ± 1.3	16.7 ± 0.3	**21.7** ± 0.9	31.3 ± 0.6
A2M	18.4 ± 1.5	18.2 ± 0.3	19.6 ± 0.8	30.6 ± 0.4
w/o PT	24.2 ± 1.1	16.9 ± 0.3	18.1 ± 0.1	33.1 ± 0.5

Table 1: Objective evaluation results. All entries are in the form of mean ± standard error.

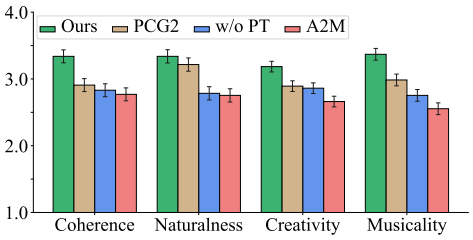


Figure 4: Subjective evaluation results.

this, we leverage CLaMP3 [28] as a cross-modal retriever. For each test audio, CLaMP3 computes the similarity between the audio and all generated symbolic piano covers. If the most similar symbolic candidate corresponds to the one generated from the audio input, we count it as a correct match. Based on this setup, we report two metrics: 1) *Top-5 Retrieval Accuracy* (Acc@5): the proportion of audio inputs for which the correct symbolic output is ranked within the top 5. Higher values indicate stronger coherence; 2) *Mean Rank*: the average rank position of the correct audio-symbolic pair across all candidates. Lower values indicate better alignment.

We conduct experiments on the POP909 test set and the Ballroom/GTZAN datasets separately. Each model is evaluated over 10 independent runs, and we report the mean and standard error. As shown in Table 1, our model consistently outperforms all baselines on POP909 by a clear margin. On the testing-only Ballroom/GTZAN datasets, it also achieves a substantially lower Mean Rank, demonstrating strong generalization across styles and genres. In comparison, the w/o PT variant surpasses PCG2 and A2M on POP909 but fails on Ballroom/GTZAN, confirming that our Stage 1 representation learning is crucial for effective style transfer and cross-modal piano arrangement. We further present an ablation study on the impact of each Stage-1 pre-training objective in Appendix D.

3.2 Subjective Evaluation

We further conduct a double-blind online listening survey to evaluate the music quality. The survey comprises 6 test pieces of varied genres drawn from the Ballroom and the GTZAN datasets. Each test piece is accompanied by 4 piano covers interpreted by our model and each baseline model. For each model, we select the best result from 3 generated samples. All samples are 16 bars long and rendered to audio using the Cakewalk TTS-1 soundfont, resulting in ~40s audio per sample. Both the order of the test pieces and the order of samples are randomized. Participants are asked to complete 3 test pieces by rating each piano cover on a 5-point Likert scale across 4 criteria: 1) *Audio-to-Symbolic Coherence*, 2) *Naturalness*, 3) *Creativity*, and 4) *Overall Musicality*.

A total of 21 participants with diverse musical backgrounds completed our survey. The average completion time is 12 minutes. Figure 4 shows the mean ratings and standard errors analyzed using within-subject ANOVA [20]. The analysis reveals significant main effects ($p < 0.05$) across all evaluation criteria. While our model performs comparably to the state-of-the-art PCG2 in Naturalness, it consistently receives higher ratings than the baselines across all criteria. A Bonferroni post-hoc test further confirms that our model significantly outperforms all baselines in Coherence and Musicality. These results align with the objective evaluation and demonstrate that our model captures music style more effectively and produces coherent, high-quality piano cover arrangements.

4 Conclusion

In this paper, we introduce a cross-modal framework for audio-to-symbolic arrangement. By repurposing the Q-Former to align audio and symbolic modalities, our model extracts and applies implicit music style using pre-trained music LMs, enabling expressive piano arrangement conditioned on both a lead sheet and an audio reference. Through a two-stage training process—combining representation learning and generative modeling—we extract stylistic features from a frozen, large audio LM and guide a symbolic LM without re-training either backbone. We conduct quantitative experiments on piano cover generation and provide qualitative demos of style transfer. Results demonstrate improved audio-to-symbolic coherence and musicality, highlighting the potential of this framework for controllable, style-aware music generation beyond explicitly labeled content.

References

[1] Hayeon Bang, Eunjin Choi, Megan Finch, Seungheon Doh, Seolhee Lee, Gyeong-Hoon Lee, and Juhan Nam. Piast: A multimodal piano dataset with audio, symbolic and text. In *Proceedings of the 3rd Workshop on NLP for Music and Audio (NLP4MusA)*, pages 5–10, 2024.

[2] Keshav Bhandari, Abhinaba Roy, Kyra Wang, Geeta Puri, Simon Colton, and Dorien Herremans. Text2midi: Generating symbolic music from captions. In *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence*, pages 23478–23486. AAAI Press, 2025.

[3] Rodrigo Castellon, Chris Donahue, and Percy Liang. Codified audio language modeling learns useful representations for music information retrieval. In *Proceedings of the 22nd International Society for Music Information Retrieval Conference, ISMIR 2021*, pages 88–96, 2021.

[4] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. In *Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023*, 2023.

[5] Christian Dittmar, Martin Pfleiderer, and Meinard Müller. Automated estimation of ride cymbal swing ratios in jazz recordings. In *Proceedings of the 16th International Society for Music Information Retrieval Conference, ISMIR 2015*, pages 271–277, 2015.

[6] Chris Donahue and Percy Liang. Sheet sage: Lead sheets from music audio. *ISMIR 2021 Late-Breaking and Demo*, 2021.

[7] Chris Donahue, John Thickstun, and Percy Liang. Melody transcription via generative pre-training. In *Proceedings of the 23rd International Society for Music Information Retrieval Conference, ISMIR 2022*, pages 485–492, 2022.

[8] Fabien Gouyon, Anssi Klapuri, Simon Dixon, M. Alonso, George Tzanetakis, C. Uhle, and Pedro Cano. An experimental comparison of audio tempo induction algorithms. *IEEE Trans. Speech Audio Process.*, 14(5):1832–1844, 2006.

[9] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022*. OpenReview.net, 2022.

[10] Yu-Siang Huang and Yi-Hsuan Yang. Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions. In *MM '20: The 28th ACM International Conference on Multimedia*, pages 1180–1188, 2020.

[11] Florian Krebs, Sebastian Böck, and Gerhard Widmer. Rhythmic pattern modeling for beat and downbeat tracking in musical audio. In *Proceedings of the 14th International Society for Music Information Retrieval Conference, ISMIR 2013*, pages 227–232, 2013.

[12] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR, 2023.

[13] Liwei Lin, Gus Xia, Junyan Jiang, and Yixiao Zhang. Content-based controls for music large language modeling. In *Proceedings of the 25th International Society for Music Information Retrieval Conference, ISMIR 2024*, pages 783–790, 2024.

[14] Liwei Lin, Gus Xia, Yixiao Zhang, and Junyan Jiang. Arrange, inpaint, and refine: Steerable long-term music audio generation and editing via content-based controls. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI 2024*, pages 7690–7698. ijcai.org, 2024.

[15] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019*. OpenReview.net, 2019.

[16] 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.

[17] Wenye Ma and Gus Xia. Exploring the internal mechanisms of music llms: A study of root and quality via probing and intervention techniques. In *ICML 2024 Workshop on Mechanistic Interpretability*, 2024.

[18] Wenye Ma, Xinyue Li, and Gus Xia. Do music llms learn symbolic concepts? a pilot study using probing and intervention. In *Audio Imagination: NeurIPS 2024 Workshop AI-Driven Speech, Music, and Sound Generation*, 2024.

[19] Ugo Marchand and Geoffroy Peeters. Swing ratio estimation. In *Proceedings of the 18th International Conference on Digital Audio Effects, DAFx-15*, pages 1–6, 2015.

[20] Henry Scheffe. *The analysis of variance*, volume 72. John Wiley & Sons, 1999.

[21] Chih-Pin Tan, Hsin Ai, Yi-Hsin Chang, Shuen-Huei Guan, and Yi-Hsuan Yang. Picogen2: Piano cover generation with transfer learning approach and weakly aligned data. In *Proceedings of the 25th International Society for Music Information Retrieval Conference, ISMIR 2024*, pages 555–562, 2024.

[22] George Tzanetakis and Perry R. Cook. Musical genre classification of audio signals. *IEEE Trans. Speech Audio Process.*, 10(5):293–302, 2002.

[23] Marcel A. Vélez Vásquez, Charlotte Pouw, John Ashley Burgoyne, and Willem H. Zuidema. Exploring the inner mechanisms of large generative music models. In *Proceedings of the 25th International Society for Music Information Retrieval Conference, ISMIR 2024*, pages 791–798, 2024.

[24] Ziyu Wang, Ke Chen, Junyan Jiang, Yiyi Zhang, Maoran Xu, Shuqi Dai, and Gus Xia. POP909: A pop-song dataset for music arrangement generation. In *Proceedings of the 21st International Society for Music Information Retrieval Conference, ISMIR 2020*, pages 38–45, 2020.

[25] Ziyu Wang, Dingsu Wang, Yixiao Zhang, and Gus Xia. Learning interpretable representation for controllable polyphonic music generation. In *Proceedings of the 21st International Society for Music Information Retrieval Conference, ISMIR 2020*, pages 662–669, 2020.

[26] Ziyu Wang, Dejing Xu, Gus Xia, and Ying Shan. Audio-to-symbolic arrangement via cross-modal music representation learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022*, pages 181–185. IEEE, 2022.

[27] Megan Wei, Michael Freeman, Chris Donahue, and Chen Sun. Do music generation models encode music theory? In *Proceedings of the 25th International Society for Music Information Retrieval Conference, ISMIR 2024*, pages 680–687, 2024.

[28] Shangda Wu, Zhancheng Guo, Ruibin Yuan, Junyan Jiang, Seungheon Doh, Gus Xia, Juhan Nam, Xiaobing Li, Feng Yu, and Maosong Sun. Clamp 3: Universal music information retrieval across unaligned modalities and unseen languages. In *Findings of the Association for Computational Linguistics, ACL 2025*, pages 2605–2625. Association for Computational Linguistics, 2025.

[29] Shih-Lun Wu, Chris Donahue, Shinji Watanabe, and Nicholas J Bryan. Music controlnet: Multiple time-varying controls for music generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:2692–2703, 2024.

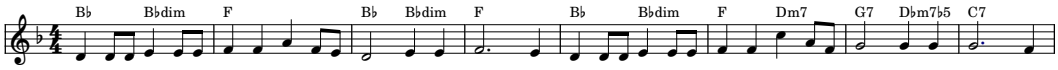
[30] Ruihan Yang, Dingsu Wang, Ziyu Wang, Tianyao Chen, Junyan Jiang, and Gus Xia. Deep music analogy via latent representation disentanglement. In *Proceedings of the 20th International Society for Music Information Retrieval Conference, ISMIR 2019*, pages 596–603, 2019.

[31] Mingliang Zeng, Xu Tan, Rui Wang, Zeqian Ju, Tao Qin, and Tie-Yan Liu. Musicbert: Symbolic music understanding with large-scale pre-training. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 791–800. Association for Computational Linguistics, 2021.

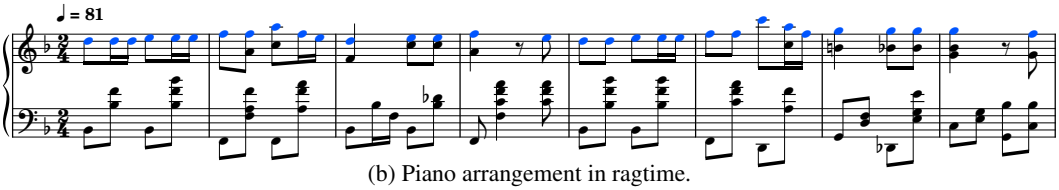
[32] Yixiao Zhang, Yukara Ikemiya, Woosung Choi, Naoki Murata, Marco A Martínez-Ramírez, Liwei Lin, Gus Xia, Wei-Hsiang Liao, Yuki Mitsufuji, and Simon Dixon. Instruct-musicgen: Unlocking text-to-music editing for music language models via instruction tuning. In *Proceedings of the 26th International Society for Music Information Retrieval Conference, ISMIR 2026*, 2026.

A Arrangement Demonstration

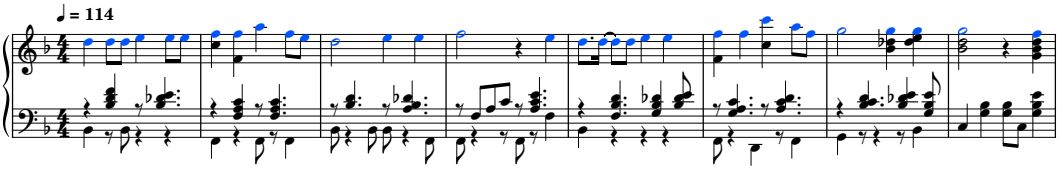
In this section, we demonstrate the performance of our audio-to-symbolic arrangement model under *freely manipulated* audio style examples. Figure 5a shows an 8-bar lead sheet excerpt from the musical *The Sound of Music*. The selected passage features harmonically rich chords, including diminished and seventh chord qualities, which present suitable complexity for arrangement experiments. Figures 5b and 5c showcase the arrangement results conditioned on two different audio examples. The 8-bar arrangement is generated using windowed sampling, wherein a 4-bar context window progresses forward every 2 bars and continues sampling conditioned upon the preceding 2 bars.



(a) The input lead sheet.



(b) Piano arrangement in ragtime.



(c) Piano arrangement in bossa nova.

Figure 5: Audio-to-symbolic arrangement for an 8-bar excerpt from *The Sound of Music*. Figures 5b and 5c are arranged based on the lead sheet in 5a and an audio reference from a ragtime and a bossa nova piece, respectively. Preserved music contents are highlighted in blue note heads. Synthesized audio is available on the demo page: `https://zhaojw1998.github.io/bossa/`.

Figure 5b shows an arrangement conditioned on the ragtime classic *The Entertainer*.[2] Following the audio recording, the arrangement's tempo is "not fast," and the piano texture distinctly adopts a ragtime rhythm, featuring steady bass notes on the downbeats and syncopated chordal accents on the upbeats. Figure 5c shows an arrangement conditioned on the bossa nova piece *The Girl from Ipanema*.[3] In this interpretation, the arrangement is characterized by a moderate tempo and distinctive left-hand syncopated patterns characteristic of the bossa nova genre.

Across both piano arrangements, while distinct music styles are effectively captured from the audio references, the theme melody and harmonic structures remain faithfully preserved. In Figure 5, we highlight melody notes preserved from the lead sheet using blue note heads.

B Datasets

Our model is trained on two dual-modal music datasets: POP909[24] and PIAST [1]. POP909 contains 1K piano cover arrangements created by professional musicians. The music genre is primarily Chinese pop while the accompanying audio features diverse band instrumentation, which can help the model learn generalizable audio representations of pop music. PIAST, on the other hand, contains 8K piano performance recordings along with symbolic transcriptions across a variety of genres, including pop, jazz, and classical. Despite the lack of band instrumentation in the piano recording, the genre diversity encourages the model to produce more expressive and stylistically varied performances. We split both datasets at song level into training (90%), validation (5%), and test (5%) sets. Each symbolic MIDI file is clipped into 4-bar segments with a 2-bar hop size and is paired with a 10s audio clip. The audio and MIDI pairs are *loosely* aligned in the center of the segment but are not necessarily synchronized at the note level. This setup prevents the model from learning low-level note-to-note transcription. Instead, it encourages extracting a general *style* representation

[2]Ragtime audio: `https://youtu.be/jKlfNfRZL9I&t=11`.

[3]Bossa nova audio: `https://youtu.be/DvA_wDOVD10&t=12`.

	PIAST		POP909	
	Acc@1 (%) ↑	**Rank** ↓	**Acc@1 (%)** ↑	**Rank** ↓
C	94.8 ± 0.4	2.6 ± 0.3	35.1 ± 0.1	5.2 ± 0.3
C+M	95.9 ± 0.3	$\mathbf{2.2} \pm 0.2$	41.6 ± 0.1	5.0 ± 0.4
C+M+G	$\mathbf{96.4} \pm 0.3$	2.3 ± 0.3	$\mathbf{44.6} \pm 0.1$	$\mathbf{4.6} \pm 0.4$

Table 2: Ablation study on the impact of individual pre-training objectives.

from the audio modality. To further encourage style abstraction, each MIDI segment is randomly transposed to all 12 keys during training.

We also test on two out-of-domain datasets: Ballroom [8, 11] and GTZAN [5, 22]. Both datasets feature audio recordings with diverse band and orchestral instrumentation, as well as fine-grained music genres such as jive and bossa nova. Since they lack paired symbolic annotations, we use them for testing only. This allows us to assess the model's generalization ability and its capacity to accommodate styles beyond pop music.

C Model Configuration and Training Details

We use MusicGen-Large [4] as our audio LM. We discard the text encoder and retain only the music decoder, a 48-layer Transformer. Audio codecs are fed to the decoder and we extract the hidden representations from the 25th layer, as prior probing studies [3, 17, 18, 23, 27] suggest that middle layers capture more musically meaningful features. This setup retains 1.7B frozen parameters from MusicGen. For symbolic music arrangement, we adopt MuseCoco-xLarge [16], a 24-layer Transformer LM pre-trained on large-scale symbolic music corpora. We remove its text-related components and keep 1.2B frozen parameters from the music decoder.

We initialize the Q-Former weights using the pre-trained MusicBERT-Base model [31]. The added cross-attention layers are randomly initialized. Following Blip-2 [12], we use $K = 32$ queries with dimension 768. The Q-Former comprises 186M parameters, including the learnable querying embeddings, which is significantly smaller than the billion-scale backbone LMs. In Stage 1, it is pre-trained in FP16 using batch size 128 for 10 epochs. The LoRA adaptor in Stage 2 has rank 16 and adds 5M parameters, and we fine-tune the model for another 5 epochs using batch size 32. Both training stages are conducted on four RTX A40 GPUs (48GB each). We use the AdamW optimizer [15] with an initial learning rate of 1e-4, a linear warm-up over the first 1k steps, and a cosine decay schedule to a final rate of 1e-5. At test time, we use top-k sampling with $k = 15$.

D Ablation Study on Pre-Training Objectives

To evaluate the contribution of each pre-training objective to cross-modal representation learning, we conduct an ablation study on the Q-Former's audio-to-symbolic retrieval performance after Stage-1 training. In this case, we repurpose the pre-trained Q-Former in Section 2.1 as a cross-modal retrieval model: for each audio clip in the test set, the Q-Former computes the similarity between the audio and all generated symbolic pieces. Similar to the setup in Section 3.1, if the most similar symbolic output corresponds to the one generated from that audio input, we count it as a correct match.

We test three configurations: contrastive loss only (**C**), contrastive + matching losses (**C+M**), and contrastive + matching + generative losses (**C+M+G**). We report *Top-1 Retrieval Accuracy* (Acc@1) and *Mean Rank* using the Q-Former on 128 randomly sampled audio-symbolic 4-bar pairs. Each experiment is repeated over 10 independent rounds, and we report the mean and standard error.

Evaluation is conducted separately on the test sets of PIAST and POP909. The former involves piano-only music, while the latter includes multi-instrumental accompaniments, requiring the model to extract style from richer audio textures. As shown in Table 2, the performance difference is relatively small on PIAST, suggesting that contrastive learning alone may suffice for simpler piano alignment. However, on POP909, we observe that both the matching and generative losses contribute meaningfully to an improved Retrieval Accuracy and a lower Mean Rank. These findings indicate that all three objectives are important for learning robust, generalizable cross-modal representations.