# Demonstrating Singing accompaniment capabilities for MuseControlLite

**Fang-Duo Tsai** **Yi-Hsuan Yang**
College of EECS, National Taiwan University, Taipei 106319, Taiwan
fundwotsai2001@gmail.com, yhyangtw@ntu.edu.tw

## Abstract

In this demo, we extend our previous work, MuseControlLite, which represents the state-of-the-art approach for controlling time-varying conditions in text-to-music models, to the task of singing accompaniment generation. Given a vocal track with MIDI and audio, MuseControlLite generates a corresponding backing track by conditioning on extracted melody, rhythm, structure, along with the local key information. This enables the system to produce musically coherent accompaniments that align with the input vocals. The demo is publicly available at: https://musecontrollite.github.io/web/.

## 1 Introduction

Text-to-music models [Borsos et al., 2023, Agostinelli et al., 2023, Copet et al., 2024, Evans et al., 2025, Tsai et al., 2024, Lee et al., 2025] have gained significant popularity over the past two years, enabling both amateurs and professionals to create music directly from text prompts. Despite their success, these models often lack precise controllability over specific musical attributes. To address this limitation, Music ControlNet [Wu et al., 2024] was the first to introduce the use of the well-known ControlNet framework [Zhang et al., 2023], allowing text-to-music models to incorporate time-varying controls. Building upon this idea, MuseControlLite [Tsai et al., 2025] further enhanced controllability with a substantially more efficient design, using a set of adapters that is 6.75 times smaller, thereby lowering the entry barrier for users. In this demo, we explore the application of MuseControlLite to the task of singing accompaniment generation (SAG).

Recent research has also investigated generating instrumental accompaniments conditioned on input vocals. SingSong [Donahue et al., 2023] employed source separation models [Kim et al., 2021] to produce aligned vocal–instrumental pairs, which were then used to train AudioLM [Borsos et al., 2023] to generate backing tracks conditioned on vocals. More recently, FastSAG [Chen et al., 2024] proposed a non-autoregressive diffusion-based approach to accelerate inference. Their method introduced semantic and prior losses to ensure rhythmic coherence between the vocals and the generated accompaniment.

However, these approaches typically require training large models from scratch. Consequently, they either suffer from slow training speed [Donahue et al., 2023] or are limited to generating short clips of only about 10 seconds. To address these challenges, we propose decomposing the vocal input into multiple time-varying conditions, enabling the model to focus on only the most relevant factors for SAG. This design provides more efficient control while preserving musical coherence.

## 2 Adapting MuseControlLite for SAG

We prepare a large dataset of Chinese pop music and apply Mel-band RoFormer [Wang et al., 2023] to separate the vocals from the backing tracks. We then extract the essential time-varying conditions for singing-accompaniment generation, which will be discussed in Section 2.1.

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

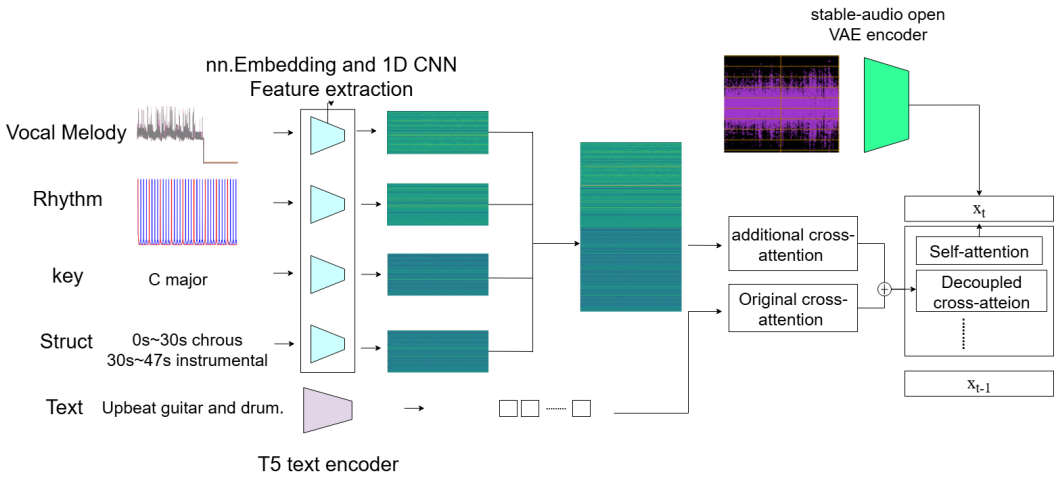


Figure 1: Model overview.

### 2.1 Condition Extraction

**Vocal Melody.** Following Hou et al. [2025], we first extract the CQT with 128 bins for both the left and right channels, and then apply an argmax operation to retain the four most prominent pitches per frame in each channel. With this vocal melody information, the model can achieve better melody harmonization.

**Structure.** Structural information provides hints for tension, intensity, and instrumentation, which should be considered in order to generate plausible accompaniment. We use All-in-one [Kim and Nam, 2023] to obtain the structural information for every song in our dataset.

**Local Key.** We use Key-CNN [Schreiber and Müller, 2019] to detect the local key of each song. The local key condition further helps the model harmonize the vocal melody more effectively.

**Beat and Downbeat Timesteps.** To synchronize rhythm, we use BeatNet [Heydari et al., 2021] to annotate beats and downbeats on the training backing tracks. During inference—when no backing track is available—we extract beats and downbeats from the vocal-track MIDI. BeatNet [Heydari et al., 2021] cannot detect beats and downbeats directly from the vocal input.

### 2.2 Training and Inference

We follow the training procedure in MuseControlLite [Tsai et al., 2025] and train the model for 7 days on a single NVIDIA RTX 3090. We drop each condition with a probability of 5%. For inference, we use 50 denoising steps to generate accompaniment for 47-second vocal tracks: the vocal audio is used to extract the melody, and the MIDI file provides the beat information. The structure and key information can be specified by the user or simply set to be the same as the vocal input.

### 2.3 Limitations

It is worth noting that our method differs slightly from other SAG approaches. Specifically, our framework requires access to the MIDI representation of the vocal input, from which we can extract beat and downbeat information. Thus, our method is more suitable for integration with singing voice synthesis (SVS) systems. In contrast, prior SAG methods operate directly on raw vocal audio without relying on MIDI.

## References

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. Audiolm: a language modeling approach to audio generation. *IEEE/ACM transactions on audio, speech, and language processing*, 31:2523–2533, 2023.

Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. Musiclm: Generating music from text. *arXiv preprint arXiv:2301.11325*, 2023.

Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation. *Advances in Neural Information Processing Systems*, 36, 2024.

Zach Evans, Julian D Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons. Stable audio open. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

Fang-Duo Tsai, Shih-Lun Wu, Haven Kim, Bo-Yu Chen, Hao-Chung Cheng, and Yi-Hsuan Yang. Audio prompt adapter: Unleashing music editing abilities for text-to-music with lightweight finetuning. *arXiv preprint arXiv:2407.16564*, 2024.

Wei-Jaw Lee, Fang-Chih Hsieh, Xuanjun Chen, Fang-Duo Tsai, and Yi-Hsuan Yang. Exploring state-space-model based language model in music generation. *arXiv preprint arXiv:2507.06674*, 2025.

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.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023.

Fang-Duo Tsai, Shih-Lun Wu, Weijaw Lee, Sheng-Ping Yang, Bo-Rui Chen, Hao-Chung Cheng, and Yi-Hsuan Yang. Musecontrollite: Multifunctional music generation with lightweight conditioners. *arXiv preprint arXiv:2506.18729*, 2025.

Chris Donahue, Antoine Caillon, Adam Roberts, Ethan Manilow, Philippe Esling, Andrea Agostinelli, Mauro Verzetti, Ian Simon, Olivier Pietquin, Neil Zeghidour, et al. Singsong: Generating musical accompaniments from singing. *arXiv preprint arXiv:2301.12662*, 2023.

Minseok Kim, Woosung Choi, Jaehwa Chung, Daewon Lee, and Soonyoung Jung. Kuielab-mdx-net: A two-stream neural network for music demixing. *arXiv preprint arXiv:2111.12203*, 2021.

Jianyi Chen, Wei Xue, Xu Tan, Zhen Ye, Qifeng Liu, and Yike Guo. Fastsag: towards fast non-autoregressive singing accompaniment generation. *arXiv preprint arXiv:2405.07682*, 2024.

Ju-Chiang Wang, Wei-Tsung Lu, and Minz Won. Mel-band roformer for music source separation. *arXiv preprint arXiv:2310.01809*, 2023.

Siyuan Hou, Shansong Liu, Ruibin Yuan, Wei Xue, Ying Shan, Mangsuo Zhao, and Chao Zhang. Editing music with melody and text: Using controlnet for diffusion transformer. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

Taejun Kim and Juhan Nam. All-in-one metrical and functional structure analysis with neighborhood attentions on demixed audio. In *2023 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pages 1–5. IEEE, 2023.

Hendrik Schreiber and Meinard Müller. Musical tempo and key estimation using convolutional neural networks with directional filters. In *Proceedings of the Sound and Music Computing Conference (SMC)*, pages 47–54, Málaga, Spain, 2019.

Mojtaba Heydari, Frank Cwitkowitz, and Zhiyao Duan. Beatnet: Crnn and particle filtering for online joint beat downbeat and meter tracking. *arXiv preprint arXiv:2108.03576*, 2021.