

ANYACCOMP: GENERALIZABLE ACCOMPANIMENT GENERATION VIA QUANTIZED MELODIC BOTTLENECK

Junan Zhang*

Yunjia Zhang*

Xueyao Zhang

Zhizheng Wu

The Chinese University of Hong Kong, Shenzhen

ABSTRACT

Singing Accompaniment Generation (SAG) is the process of generating instrumental music for a given clean vocal input. However, existing SAG techniques use source-separated vocals as input and overfit to separation artifacts. This creates a critical train-test mismatch, leading to failure on clean, real-world vocal inputs. We introduce ANYACCOMP, a framework that resolves this by decoupling accompaniment generation from source-dependent artifacts. ANYACCOMP first employs a quantized melodic bottleneck, using a chromagram and a VQ-VAE to extract a discrete and timbre-invariant representation of the core melody. A subsequent flow-matching model then generates the accompaniment conditioned on these robust codes. Experiments show ANYACCOMP achieves competitive performance on separated-vocal benchmarks while significantly outperforming baselines on generalization test sets of clean studio vocals and, notably, solo instrumental tracks. This demonstrates a qualitative leap in generalization, enabling robust accompaniment for instruments—a task where existing models completely fail—and paving the way for more versatile music co-creation tools. Demo audio and code: <https://anyaccomp.github.io/>

Index Terms— Music Generation, Singing Accompaniment Generation, Vector Quantization, Flow Matching

1. INTRODUCTION

Singing Accompaniment Generation (SAG) is a crucial task in music creation, aiming to generate instrumental music that harmonically and rhythmically complements a given vocal melody [1, 2]. High-fidelity SAG models hold immense potential to revolutionize the music creation process, offering powerful co-creation tools for artists, enabling producers to rapidly prototype ideas, and empowering amateurs to bring their musical visions to life.

A common training paradigm for state-of-the-art SAG models [1–3] relies on vocal data extracted by Music Source Separation (MSS) algorithms [4, 5]. The vocal conditioning signal is typically a direct acoustic representation, such as the mel spectrogram used in FastSAG [2], or features from powerful Self-Supervised Learning (SSL) models, an approach

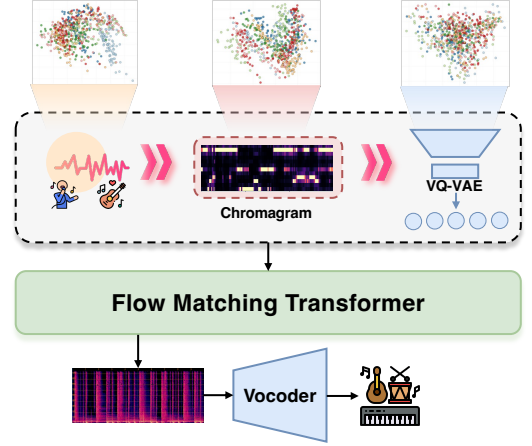


Fig. 1: An overview of ANYACCOMP. The process involves two main stages: (1) The input audio is processed through a quantized melodic bottleneck, where a VQ-VAE encodes its chromagram into a sequence of discrete tokens. (2) A Flow Matching Transformer generates a mel-spectrogram conditioned on these discrete tokens, which is subsequently synthesized into the final accompaniment audio by a vocoder.

pioneered by SingSong [1] with a custom w2v-BERT [6]. However, both mel spectrograms and general-purpose SSL features are designed to preserve rich acoustic details. Consequently, when trained on imperfectly separated data, they invariably learn an artifactual correlation, associating valid accompaniment generation with the presence of separation artifacts. While previous works [1, 2] attempt to mitigate this by injecting white noise, this can introduce new, unrealistic artifacts. This leads to a critical **train-test mismatch**: in real-world scenarios where recordings are clean, these models falter because they lack the very “cues” they were trained to expect. This dependency on imperfect data significantly limits the robustness and practical applicability of current SAG systems.

In this paper, we view the SAG task as a conditional generation problem, arguing the mismatch stems from the conditional representation’s sensitivity to source-specific characteristics. We therefore propose a representation designed to be robust by satisfying two key properties: **timbre invariance** to filter out source-dependent textures such as timbre or separation artifacts, while **melodic clusterability** preserves the essential melodic content in a compact and structured form. This pro-

*Equal Contribution.

vides a unified condition that resolves the train-test mismatch, enabling a model trained solely on vocal data to robustly generalize to clean vocals and even instrumental tracks.

To this end, we propose ANYACCOMP, a novel two-stage framework for robust accompaniment generation. The first stage of ANYACCOMP is a bottleneck module that extracts a generalized melodic representation by using a Vector-Quantized Variational Autoencoder (VQ-VAE) [7] to quantize the input audio’s timbre-invariant chromagram into a sequence of discrete codes. This process effectively isolates the core melodic content, creating a representation that is robust to source-dependent artifacts. The efficacy of this representation is empirically demonstrated in Figure 2, which shows our codes achieving significantly clearer melodic clusterability and timbre invariance compared to traditional mel spectrograms.

In the second stage, a Flow-Matching Transformer [8] generates the accompaniment conditioned on these robust and generalized representations. By decoupling the generation process from artifact-preserving representations, ANYACCOMP avoids the correlations with separation artifacts that limit existing methods. Our experiments validate this approach: ANYACCOMP not only performs competitively on in-domain separated-vocal benchmarks but also significantly outperforms baselines on generalization test sets with clean studio vocals and, notably, solo instrumental tracks. This success demonstrates that our method effectively resolves the critical train-test mismatch. The main contributions of this work are as follows:

- We introduce a quantized melodic representation via a chromagram and VQ-VAE bottleneck, which is timbre-invariant and robust to separation artifacts.
- We propose ANYACCOMP, a two-stage framework that decouples accompaniment generation from artifact-prone inputs via a Flow-Matching Transformer, and demonstrate its effectiveness in resolving the train-test mismatch by outperforming baselines on clean studio vocals and solo instrumental tracks.

2. ANYACCOMP FRAMEWORK

2.1. Quantized Melodic Bottleneck

The core of our approach is to learn an intermediate representation that captures essential melodic content while discarding irrelevant timbral information and source separation artifacts. To obtain this generalized representation, we employ a Vector-Quantized Variational Autoencoder (VQ-VAE) that operates on a 50 Hz Chromagram input $\mathbf{x}$. The model’s encoder maps the chromagram to a continuous latent representation $\mathbf{z}_e(\mathbf{x})$, which is then quantized to its nearest neighbor $\mathbf{e}_k$ from a learned codebook $\mathbf{E}$ via $k = \arg \min_j \|\mathbf{z}_e(\mathbf{x}) - \mathbf{e}_j\|_2$. A decoder then reconstructs the chromagram $\hat{\mathbf{x}}$ from the resulting quantized sequence $\mathbf{z}_q(\mathbf{x})$. The model is trained to minimize a joint loss function:

$$\mathcal{L} = \|\mathbf{x} - \hat{\mathbf{x}}\|_2^2 + \beta \|\mathbf{z}_e(\mathbf{x}) - \mathbf{z}_q(\mathbf{x})\|_2^2 \quad (1)$$

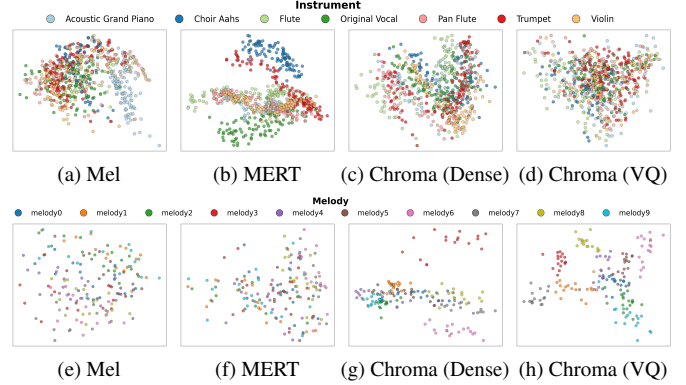


Fig. 2: Visualization of different representations. The top row evaluates **timbre invariance**, where better representations show instrument colors that are more thoroughly intermingled. The bottom row assesses **melodic clusterability**, where better representations form tighter, more distinct clusters for each melody color.

This quantization process forces the model to capture crucial melodic features while discarding artifacts, creating a robust bottleneck for the generation stage.

To evaluate our representation’s effectiveness, we conduct a visualization experiment using melodies from the M4Singer dataset [9]. For each melody, we use the original vocal utterance and augment it by synthesizing several instrumental versions from its corresponding MIDI data using `pretty_midi`¹. This creates a controlled set where the melody is constant while the timbre varies. From this combined collection, we extract and compare four representations: the standard mel spectrogram, features from the pre-trained MERT model [10], a dense chromagram, and our VQ-quantized chromagram. Figure 2 visualizes their PCA-reduced embeddings to evaluate two key properties.

Timbre invariance. A robust representation must be agnostic to the source timbre, a property crucial for generalization. As visualized in the top row of Figure 2, representations that preserve rich acoustic features fail this test. The traditional mel spectrogram (Figure 2a) exhibits strong clustering by instrument type. This effect is even more pronounced with the MERT representation (Figure 2b), as its pre-training objective is designed to capture fine-grained acoustic characteristics. The dense chromagram (Figure 2c) significantly mitigates this bias by breaking up these obvious clusters. However, it is our final vector-quantized chromagram representation (Figure 2d) that achieves the ideal result: a feature space where points from all instruments are thoroughly blended, demonstrating a successful decoupling of melodic content from source timbre.

Melodic clusterability. An effective representation must map identical melodies to compact regions in the feature space, irrespective of the instrument. The bottom row of Figure 2 illustrates a clear progression. Both the mel spectrogram (Fig-

¹<https://github.com/craffel/pretty-midi>

ure 2e) and the MERT representation (Figure 2f) perform poorly, failing to form coherent melodic clusters. The dense chromagram (Figure 2g) shows a marked improvement, with distinct melodic clusters becoming visible. Our final vector-quantized chromagram representation (Figure 2h) further refines this by producing exceptionally tight and well-separated clusters, demonstrating that quantization not only preserves but also enhances the melodic structure, creating a highly robust and discrete condition for the generative model.

2.2. Flow-Matching Accompaniment Generation

In the second stage, a Flow-Matching (FM) Transformer [8] generates the accompaniment conditioned on the discrete melodic codes. Crucially, this strategy decouples the generation process from input source artifacts, ensuring the output is based purely on the musical melody and rhythm.

We formulate accompaniment generation as a conditional flow-matching problem, where the goal is to learn a continuous-time probability path $p_t(\mathbf{x})$ that transforms a simple prior distribution into the target data distribution. Let $\mathbf{x}_1$ represent the target accompaniment mel spectrogram. The path begins from $\mathbf{x}_0 \sim \mathcal{N}(0, \mathbf{I})$, a sample from a standard Gaussian distribution. We define the trajectory between these two points as:

$$\mathbf{x}_t = (1 - (1 - \sigma)t)\mathbf{x}_0 + t\mathbf{x}_1 \quad (2)$$

where $t \in [0, 1]$ is the timestep and σ is a small constant (e.g., 10^{-5}). During training, the timestep t is sampled from a cosine schedule by first drawing $t' \sim U[0, 1]$ and then applying $t = 1 - \cos(t' \cdot \frac{\pi}{2})$ to concentrate sampling density near the beginning of the trajectory.

The model is trained to predict the vector field $\mathbf{v}_t$ that generates this path. The ground-truth velocity is the time derivative of $\mathbf{x}_t$:

$$\mathbf{v}_t = \frac{d\mathbf{x}_t}{dt} = \mathbf{x}_1 - (1 - \sigma)\mathbf{x}_0 \quad (3)$$

Our model f_θ , conditioned on the discrete melodic codes $\mathbf{c}$, learns to approximate this velocity field. The training objective is a simple mean squared error loss:

$$\mathcal{L}_{FM}(\theta) = \mathbb{E}_{t, \mathbf{x}_1, \mathbf{x}_0, \mathbf{c}} \|f_\theta(\mathbf{x}_t, t, \mathbf{c}) - \mathbf{v}_t\|^2 \quad (4)$$

We add a Representation Alignment (REPA) loss [11], which aligns an intermediate FM Transformer layer with representations from a pre-trained music model [10]. The final objective is the weighted sum of the two losses: $\mathcal{L} = \mathcal{L}_{FM} + \lambda\mathcal{L}_{REPA}$.

During inference, we generate the accompaniment mel spectrogram $\mathbf{x}_1$ by starting from a random noise sample $\mathbf{x}_0$ and integrating the predicted velocity $\hat{\mathbf{v}}_t = f_\theta(\mathbf{x}_t, t, \mathbf{c})$ from $t = 0$ to $t = 1$ using the forward Euler method: $\mathbf{x}_{t+h} = \mathbf{x}_t + h\hat{\mathbf{v}}_t$, where h is the step size computed as the inverse of the total

timesteps dedicated for integration. To leverage classifier-free guidance (CFG) [12], we randomly drop the melodic condition $\mathbf{c}$ with a probability of 0.1 during the training process.

3. EXPERIMENTS

3.1. Experimental Setup

Training Data. We prepare 8k hours of paired singing voice–accompaniment data for training ANYACCOMP, following the SingNet pipeline [13]. The data is sourced from in-the-wild songs on the Internet, separated into vocals and accompaniment, and then sliced into clips ranging from 3 s to 30 s. All audio is processed at a 24 kHz sampling rate.

Implementation Details. Our VQ-VAE (44M params), adapted from Amphion [14, 15], quantizes a 24-bin chromagram into a 50 Hz sequence with a 512-entry codebook, and is trained for 0.5M steps with a batch size of 200 seconds. The Flow-Matching (FM) Transformer, based on Vevo [16, 17], consists of 10 LLaMA decoder layers with a hidden dimension of 1024, totaling 220M parameters. We fine-tune the vocoder from Vevo on our music data for use in our system. FM models are trained for 1M steps with a per-GPU batch size of 100 seconds, incorporating REPA loss with MERT-330M [10] as the alignment target at layer 4 of $\lambda = 0.5$ weight. All models are optimized with AdamW [18] (learning rate $1e-4$, 32k warmup steps) and trained on a single GPU. During inference, we use 50 sampling steps with a CFG scale of 3.

Compared Methods. To evaluate our melodic bottleneck, we compare ANYACCOMP with FastSAG [2], a state-of-the-art non-autoregressive model using vocal mel-spectrograms, and two controlled variants that share the FM Transformer architecture of ANYACCOMP: FM-Mel, conditioned on mel spectrograms with added white noise (15-20 dB SNR), and FM-Chroma, conditioned on raw 24-bin chromagrams. ANYACCOMP instead uses quantized chromagram codes.

Evaluation Datasets. We evaluate our models on three distinct 10-second clip datasets to test different capabilities. For in-domain performance, we use 3,000 separated vocal-accompaniment pairs from the YuE data [19]. To assess generalization to artifact-free audio, we use 2,777 clean vocal stems from the MUSDB18 test set [20]. Finally, to push the boundaries of generalization, we evaluate on 2,500 solo instrumental tracks from MoisesDB [21].

Evaluation Metrics. We assess accompaniment with objective and subjective metrics. Objective measures include Fréchet Audio Distance (FAD) [22] for distributional similarity, Accompaniment Prompt Adherence (APA) [23] for condition alignment, and audibox-aesthetics [24] scores on Content Enjoyment (CE), Content Usefulness (CU), Production Complexity (PC), and Production Quality (PQ). For subjective evaluation, we ran a MOS test on 20 samples per test set, where listeners rated overall quality and coherency on a 1–5 scale.

Table 1: Objective evaluation results showing that ANYACOMP is competitive on in-domain separated vocals (YuE) while strongly outperforming baselines on generalization to clean vocals (MUSDB18) and instruments (MoisesDB).

Model	APA $\uparrow$	FAD $\downarrow$	CE $\uparrow$	CU $\uparrow$	PQ $\uparrow$	PC $-$
YuE (In-domain Separated Vocal)						
Ground Truth	-	-	7.270	7.784	7.734	5.752
FastSAG	0.444	0.598	6.351	6.821	6.814	6.321
FM-Mel	0.806	0.416	6.964	7.725	7.758	5.614
FM-Chroma	0.633	0.418	7.151	7.801	7.909	5.436
ANYACOMP	0.713	0.414	7.283	7.903	7.989	5.742
MUSDB18 (Clean Vocal)						
Ground Truth	-	-	7.164	7.616	7.485	5.957
FastSAG	0.000	1.115	4.853	5.789	6.315	5.778
FM-Mel	0.167	0.999	5.202	6.616	6.841	4.090
FM-Chroma	0.704	0.798	7.017	7.598	7.744	5.104
ANYACOMP	0.710	0.788	7.277	7.804	7.891	5.498
MoisesDB (Instruments)						
Ground Truth	-	-	7.236	7.791	7.778	5.694
FastSAG	0.000	0.904	5.966	6.507	6.696	5.952
FM-Mel	0.000	0.936	5.424	6.923	7.151	3.804
FM-Chroma	0.157	0.849	6.308	7.377	7.508	4.110
ANYACOMP	0.203	0.890	6.660	7.581	7.581	4.798

3.2. Results

The objective results in Table 1 confirm our bottleneck representation’s pivotal role in robust generalization. While ANYACOMP is competitive on the in-domain YuE dataset, its true advantage emerges where the train-test mismatch becomes critical. On clean vocals (MUSDB18), mel-spectrogram-conditioned models (FastSAG, FM-Mel) suffer a catastrophic collapse, with APA scores drop to 0. This score signifies a complete failure in conditional generation, as the resulting accompaniment has no correlation to the input melody—a direct consequence of severe overfitting to separation artifacts. This failure is even more pronounced on the more out-of-domain instrumental tracks from MoisesDB.

In stark contrast, ANYACOMP, leveraging by its quantized melodic bottleneck, maintains high coherence and quality across all generalization sets. This demonstrates its superior generalization capability (validated by the timbre invariance and melodic clusterability shown in Figure 2), enabling it to overcome the train-test mismatch and robustly generate accompaniment for entirely new timbres where other models completely fail. We also note that Production Complexity (PC) measures textural complexity, not perceptual quality, and is thus not a “higher-is-better” metric.

As illustrated in the case study of Figure 3, FM-Mel suffers from severe spectral leakage, directly copying artifacts from the input vocal. In contrast, ANYACOMP generates a coherent

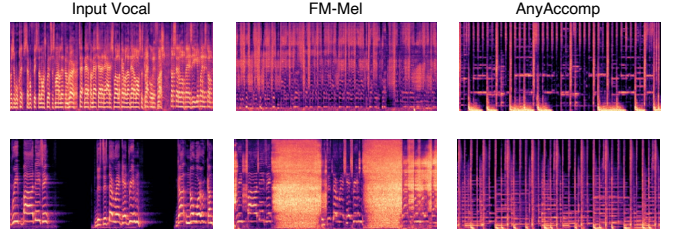


Fig. 3: A case study on a clean MUSDB18 vocal. FM-Mel’s output exhibits severe **spectral leakage** from the input, a sign of overfitting to source-separation artifacts. In contrast, ANYACOMP generates a coherent instrumental accompaniment.

Table 2: Subjective evaluation results on the three test sets. Quality refers to the overall accompaniment quality, while Coherency measures how well the accompaniment matches the input.

Model	Quality $\uparrow$	Coherency $\uparrow$
YuE (In-domain Separated Vocal)		
Ground Truth	3.92	3.88
FastSAG	1.98	1.82
ANYACOMP	3.12	3.05
MUSDB18 (Clean Vocal)		
Ground Truth	3.65	3.48
FastSAG	1.73	1.48
ANYACOMP	3.23	2.75
MoisesDB (Instruments)		
Ground Truth	4.05	4.08
FastSAG	1.62	1.52
ANYACOMP	3.00	2.70

accompaniment, highlighting its robustness against overfitting.

The subjective listening result (Table 2) confirms our objective findings. Listeners rated ANYACOMP significantly higher than FastSAG in both Quality and Coherency across all datasets. Crucially, ANYACOMP maintains high scores on the challenging MUSDB18 and MoisesDB generalization sets, validating that our objective improvements translate to a superior and more robust listening experience.

4. CONCLUSION

In this work, we present ANYACOMP, a framework that resolves the critical train-test mismatch in SAG models by using a quantized chromagram bottleneck to decouple the generation process from source-separation artifacts and timbre variations. Our model achieves competitive in-domain performance while drastically outperforming baselines when generalizing to clean vocals and unseen solo instruments. Future work will focus on optimizing bottleneck parameters (e.g., frame rate and vocabulary size), and exploring alternative representations like the Constant-Q Transform (CQT).

5. REFERENCES

- [1] Chris Donahue, Antoine Caillon, Adam Roberts, Ethan Manilow, Philippe Esling, Andrea Agostinelli, Mauro Verzett, Ian Simon, Olivier Pietquin, Neil Zeghidour, et al., “Singsong: Generating musical accompaniments from singing,” *arXiv preprint arXiv:2301.12662*, 2023.
- [2] Jianyi Chen, Wei Xue, Xu Tan, Zhen Ye, Qifeng Liu, and Yike Guo, “FastSAG: towards fast non-autoregressive singing accompaniment generation,” in *Proc. IJCAI*, 2024.
- [3] Junmin Gong, Sean Zhao, Sen Wang, Shengyuan Xu, and Joe Guo, “Ace-step: A step towards music generation foundation model,” *arXiv preprint arXiv:2506.00045*, 2025.
- [4] Yi Luo and Jianwei Yu, “Music source separation with band-split RNN,” *IEEE/ACM TASLP*, vol. 31, pp. 1893–1901, 2023.
- [5] Wei-Tsung Lu, Ju-Chiang Wang, Qiuqiang Kong, and Yun-Ning Hung, “Music source separation with band-split rope transformer,” in *Proc. ICASSP*, 2024.
- [6] Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu, “W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training,” in *IEEE ASRU*, 2021.
- [7] Aaron Van Den Oord, Oriol Vinyals, et al., “Neural discrete representation learning,” *NeurIPS*, vol. 30, 2017.
- [8] Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le, “Flow matching for generative modeling,” *arXiv preprint arXiv:2210.02747*, 2022.
- [9] Lichao Zhang, Ruiqi Li, Shoutong Wang, Liqun Deng, Jinglin Liu, Yi Ren, Jinzheng He, Rongjie Huang, Jieming Zhu, Xiao Chen, et al., “M4singer: A multi-style, multi-singer and musical score provided mandarin singing corpus,” *NeurIPS*, 2022.
- [10] Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Xingran Chen, Hanzhi Yin, Chenghao Xiao, Chenghua Lin, Anton Ragni, Emmanouil Benetos, et al., “MERT: Acoustic music understanding model with large-scale self-supervised training,” in *ICLR*, 2024.
- [11] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie, “Representation alignment for generation: Training diffusion transformers is easier than you think,” in *ICLR*, 2025.
- [12] Jonathan Ho and Tim Salimans, “Classifier-free diffusion guidance,” *arXiv preprint arXiv:2207.12598*, 2022.
- [13] Yicheng Gu, Chaoren Wang, Junan Zhang, Xueyao Zhang, Zihao Fang, Haorui He, and Zhizheng Wu, “Singnet: Towards a large-scale, diverse, and in-the-wild singing voice dataset,” *arXiv preprint arXiv:2505.09325*, 2025.
- [14] Xueyao Zhang, Liumeng Xue, Yicheng Gu, Yuancheng Wang, Jiaqi Li, Haorui He, Chaoren Wang, Ting Song, Xi Chen, Zihao Fang, Haopeng Chen, Junan Zhang, Tze Ying Tang, Lexiao Zou, Mingxuan Wang, Jun Han, Kai Chen, Haizhou Li, and Zhizheng Wu, “Amphion: An open-source audio, music and speech generation toolkit,” in *Proc. IEEE SLT Workshop*, 2024.
- [15] Jiaqi Li, Xueyao Zhang, Yuancheng Wang, Haorui He, Chaoren Wang, Li Wang, Huan Liao, Junyi Ao, Zeyu Xie, Yiqiao Huang, Junan Zhang, and Zhizheng Wu, “Overview of the Amphion Toolkit (v0.2),” *arXiv preprint arXiv:2501.15442*, 2025.
- [16] Xueyao Zhang, Xiaohui Zhang, Kainan Peng, Zhenyu Tang, Vimal Manohar, Yingru Liu, Jeff Hwang, Dangna Li, Yuhao Wang, Julian Chan, et al., “Vevo: Controllable zero-shot voice imitation with self-supervised disentanglement,” in *ICLR*, 2025.
- [17] Xueyao Zhang, Junan Zhang, Yuancheng Wang, Chaoren Wang, Yuanzhe Chen, Dongya Jia, Zhuo Chen, and Zhizheng Wu, “Vevo2: Bridging controllable speech and singing voice generation via unified prosody learning,” *arXiv preprint arXiv:2508.16332*, 2025.
- [18] Ilya Loshchilov and Frank Hutter, “Decoupled weight decay regularization,” in *ICLR*, 2019.
- [19] Ruibin Yuan, Hanfeng Lin, Shuyue Guo, Ge Zhang, Jiahao Pan, Yongyi Zang, Haohe Liu, Yiming Liang, Wenye Ma, Xingjian Du, et al., “Yue: Scaling open foundation models for long-form music generation,” *arXiv preprint arXiv:2503.08638*, 2025.
- [20] Zafar Rafii, Antoine Liutkus, and Fabian-Robert Stöter, “The MUSDB18 corpus for music separation,” 2017.
- [21] Igor Pereira, Felipe Araújo, Filip Korzeniowski, and Richard Vogl, “MoisesDB: A dataset for source separation beyond 4-stems,” *arXiv preprint arXiv:2307.15913*, 2023.
- [22] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi, “Fréchet audio distance: A metric for evaluating music enhancement algorithms,” *arXiv preprint arXiv:1812.08466*, 2018.
- [23] Maarten Grachten and Javier Nistal, “Accompaniment prompt adherence: A measure for evaluating music accompaniment systems,” in *Proc. ICASSP*, 2025.
- [24] Andros Tjandra, Yi-Chiao Wu, Baishan Guo, John Hoffman, Brian Ellis, Apoorv Vyas, Bowen Shi, Sanyuan Chen, Matt Le, Nick Zacharov, et al., “Meta Audiobox Aesthetics: Unified automatic quality assessment for speech, music, and sound,” *arXiv preprint arXiv:2502.05139*, 2025.