Modeling Music as a Time-Frequency Image: A 2D Tokenizer for Music Generation

1st Yuqing Cheng[1,*,†], 2nd Xingyu Ma[2,*], 3rd Guochen Yu[2], 4th Xiaotao Gu[2]
[1]*Department of Music AI and Information Technology, Central Conservatory of Music*, Beijing, China
[2]*Zhipu AI*, Beijing, China
chengyuqing@mail.ccom.edu.cn, fujindemi@gmail.com

***Abstract*—Autoregressive music generation depends strongly on the audio tokenizer. Existing high-fidelity codecs often use residual multi-codebook quantization, which preserves reconstruction quality but complicates language modeling after sequence flattening, as the residual hierarchy imposes strong sequential dependencies and can amplify error accumulation. We propose BandTok, a generation-oriented 2D Mel-spectrogram tokenizer that represents each frame with Mel-frequency band tokens from a single shared codebook. This design yields a physically interpretable time-frequency token grid with a more independent token structure, making it better suited for autoregressive modeling. BandTok improves reconstruction with a multi-scale PatchGAN objective and EMA codebook updates. We further introduce an autoregressive language model with 2D Rotary Position Embedding (2D RoPE) to preserve temporal and frequency-band structure during generation. Experiments show that BandTok improves over residual-codebook tokenizers and achieves strong results in a data-limited setting. The source code and generation demos for this work are publicly available.[1]**

***Index Terms*—Music, tokenization, spectrogram, audio coding, large language models.**

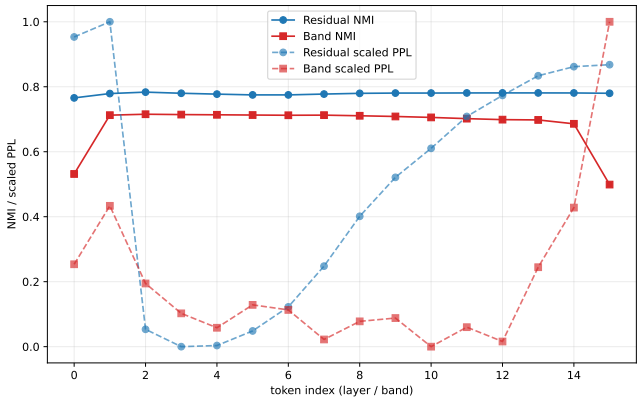

Fig. 1. Comparison between residual and band-wise tokens. Normalized mutual information (NMI) and language-model perplexity (PPL) are used to analyze token dependence and autoregressive prediction difficulty, respectively. Compared with residual tokens, band-wise tokens exhibit lower inter-token dependence and a more balanced PPL profile.

I. Introduction

Recent advances in music generation have been driven by diffusion-based generation and autoregressive token modeling. Autoregressive approaches are attractive because they leverage the scalability of language models (LMs), but their effectiveness depends critically on the audio tokenizer that converts waveforms into discrete tokens. For generation-oriented tokenization, the tokenizer must jointly satisfy high reconstruction fidelity and LM-friendly token organization. These factors determine the acoustic upper bound, sequence predictability, and error propagation behavior of autoregressive music generation.

High-fidelity neural audio codecs [1]–[3] commonly employ Residual Vector Quantization (RVQ), where multiple codebooks progressively refine reconstruction. Although RVQ preserves fine acoustic details, its residual-layer structure complicates autoregressive modeling. When multi-codebook tokens are flattened into a sequence, the LM must predict along a residual refinement hierarchy, in which later codebooks encode increasingly fine residual corrections conditioned on earlier ones. Thus, early prediction errors can propagate to later codebooks, degrading high-level token prediction and accumulating artifacts. Existing methods, including semantic-to-acoustic pipelines [4], [5], delayed codebook prediction [6], and dual-autoregressive modeling [7], mainly address this burden at the modeling stage while retaining the residual-codebook geometry. Improving token independence has been shown to facilitate downstream LM training through independence-promoting tokenizer objectives [8].

Spectral single-codebook codecs such as MelCap [9] and UniSRCodec [10] show that Mel-spectrogram-based two-dimensional tokenization can achieve compact and high-fidelity reconstruction. However, these methods are primarily evaluated as compression systems, leaving unclear whether such spectral token geometry is suitable for autoregressive music generation. In particular, a generation-oriented tokenizer must not only reconstruct well, but also produce token sequences that are stable and predictable for generation.

We propose BandTok, a two-dimensional Mel-spectrogram tokenizer for autoregressive music generation. BandTok represents each frame with low-to-high Mel-frequency band tokens using a shared codebook. After flattening the time-frequency grid for LM training, the within-frame order follows spectral

*Equal contribution. †Work done during an internship at Zhipu AI.

[1]https://github.com/xiaolubuhuizhuzhou/Bandtok

bands rather than residual refinement layers. Unlike RVQ, later tokens do not explicitly encode residual corrections conditioned on earlier codebooks, which reduces residual-hierarchy dependence and yields more stable autoregressive targets, as supported by Figure 1. We further use two-dimensional Rotary Position Embedding (2D RoPE) to preserve temporal and frequency-band position information after flattening.

To improve reconstruction fidelity and training stability, BandTok adopts a MelCap-style architecture with a multi-scale PatchGAN [11] discriminator and exponential moving average (EMA) codebook updates. The discriminator encourages perceptually important spectral detail reconstruction, while EMA stabilizes large-codebook training. Together, these designs balance high-fidelity reconstruction with LM-friendly token geometry.

Our contributions are summarized as follows:

1. We develop BandTok, a generation-oriented two-dimensional Mel-spectrogram tokenizer for autoregressive music generation. By organizing tokens along Mel-frequency bands rather than residual codebook layers, BandTok provides a physically interpretable and LM-friendly token geometry that reduces residual-chain error propagation.

2. We improve the reconstruction fidelity and training stability of spectral tokenization for music. With a MelCap-style architecture, a multi-scale PatchGAN discriminator, and EMA codebook updates, BandTok enhances high-frequency details and stabilizes large-codebook training, achieving superior reconstruction quality over waveform-domain tokenizers under comparable low-bitrate settings.

3. We introduce an autoregressive music generation framework over flattened time-frequency tokens. By incorporating 2D RoPE, the LM preserves temporal and frequency-band positional structure after flattening. Experiments show that, under comparable reconstruction quality, BandTok improves objective and subjective generation quality over alternative tokenizers and achieves stronger music generation performance under academic-scale data training.

II. Related Works

A. Autoregressive Music Generation

Autoregressive music generation relies on audio tokenizers that convert waveforms into discrete token sequences. Hierarchical systems such as AudioLM [4] and MusicLM [5] decompose generation into semantic and acoustic stages, where semantic tokens capture long-range structure and acoustic tokens reconstruct waveform-level details. However, these pipelines depend heavily on pretrained semantic representations.

MusicGen [6] simplifies this design by directly modeling EnCodec [2] tokens with a single autoregressive Transformer and delayed codebook prediction. UniAudio [7] further uses multi-scale Transformers and separate language models for coarse and fine tokens to handle token hierarchy. These designs improve the modeling of residual multi-codebook tokens, but they still inherit the strong inter-codebook dependence induced by residual quantization. This motivates us to revisit the tokenizer geometry itself as the interface for autoregressive music generation.

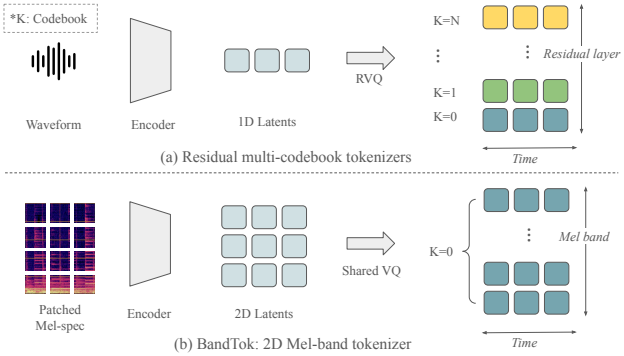

Fig. 2. Comparison between RVQ tokenizers and BandTok. Figure (a) shows RVQ-based audio tokenizers, where each VQ layer quantizes the residual from the previous layer. Figure (b) shows BandTok, which patchifies the Mel spectrogram into 2D latents and quantizes them with a single shared codebook. Its vertical axis corresponds to Mel-frequency bands.

B. Audio Tokenization

Neural audio tokenizers differ in both representation domain and token geometry. Waveform-domain codecs [1]–[3], [12] achieve high-fidelity reconstruction by directly encoding waveforms, typically with residual or multi-codebook quantization. However, this structure introduces a residual codebook axis for downstream LMs, making autoregressive prediction sensitive to inter-codebook dependence and error propagation.

Spectral-domain codecs [13]–[15] instead tokenize time-frequency representations, showing that Mel- or STFT-based representations can support high-quality audio reconstruction. However, their token streams are still primarily organized as one-dimensional or residual-codebook sequences. More recent two-dimensional spectral tokenizers, including MelCap [9] and UniSRCodec [10], exploit the image-like structure of Mel spectrograms and demonstrate strong reconstruction quality. Yet these methods are mainly evaluated as codecs, leaving the role of two-dimensional token geometry in autoregressive music generation underexplored.

III. Method

We present BandTok, a generation-oriented Mel-spectrogram tokenizer, together with an autoregressive language model over flattened time-frequency tokens. Our method is designed around two goals: improving spectral reconstruction fidelity and preserving two-dimensional token geometry for autoregressive music generation.

A. BandTok Tokenizer

For each 44.1 kHz waveform, we compute a log-Mel spectrogram $\mathbf{X} \in \mathbb{R}^{N\times 1\times T\times F}$ with $F = 128$ Mel bins, using a 2048-sample STFT window and a 512-sample hop. BandTok first applies 2D Haar [16] patchification with patch size $p = 2$, decomposing the spectrogram into LL, LH, HL, and HH subbands to retain both coarse spectral structure and local high-frequency details. A Cosmos-style [17] encoder maps the

patched spectrogram to a latent grid $\mathbf{Z}_e \in \mathbb{R}^{N\times C\times T'\times F'}$, downsampling by $8\times$ along both time and frequency. This yields an audio-token frame rate of approximately 10.7 Hz and $F' = 16$ frequency-band positions.

As shown in Figure 2, BandTok uses a single 8192-entry codebook to quantize the latent grid into a discrete two-dimensional token grid. The codebook is updated with EMA statistics instead of an explicit codebook loss, improving large-codebook stability and reducing noisy updates for rarely selected codes, while a standard commitment loss regularizes the encoder. The quantized grid is decoded back into a Mel spectrogram and converted to waveform audio using a pretrained BigVGAN-v2 vocoder [18].

B. Reconstruction Objective

To improve high-frequency reconstruction, we introduce a multi-scale PatchGAN discriminator on Mel spectrograms. Each discriminator operates on a different spectrogram resolution obtained through linear interpolation. This design encourages realistic local time-frequency details across multiple scales, which is important for preserving musical texture and high-frequency content.

The tokenizer is trained with a weighted combination of reconstruction, perceptual, adversarial, feature-matching, and commitment losses:

$$\begin{aligned}\mathcal{L}_{\text{BandTok}} =& \lambda_{\text{rec}}\mathcal{L}_{\text{rec}} + \lambda_{\text{perc}}\mathcal{L}_{\text{perc}} + \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} \\ &+ \lambda_{\text{fm}}\mathcal{L}_{\text{fm}} + \lambda_{\text{commit}}\mathcal{L}_{\text{commit}}.\end{aligned} \tag{1}$$

Here, $\mathcal{L}_{\text{rec}}$ denotes the L1 Mel-spectrogram reconstruction loss, $\mathcal{L}_{\text{perc}}$ the VGG-based perceptual loss [19], $\mathcal{L}_{\text{adv}}$ the generator-side adversarial loss, $\mathcal{L}_{\text{fm}}$ the discriminator feature-matching loss, and $\mathcal{L}_{\text{commit}}$ the VQ commitment loss. We set the corresponding weights to $\lambda_{\text{rec}} = 5.0$, $\lambda_{\text{perc}} = 1.0$, $\lambda_{\text{adv}} = 1.0$, $\lambda_{\text{fm}} = 5.0$, and $\lambda_{\text{commit}} = 2.5$.

C. Autoregressive Modeling with 2D RoPE

Applying standard 1D Rotary Position Embedding (RoPE) to the flattened sequence creates a mismatch between sequence order and spectrogram geometry. In particular, tokens from the same frequency band in adjacent frames are separated by all frequency-band positions within a frame, as shown in Figure 3. This weakens the local time-frequency inductive bias and requires the model to infer the original two-dimensional structure implicitly.

To address this issue, we adopt Interleaved-MRoPE from Qwen3-VL [20]. The attention-head dimension is split into token, time, and frequency-band components, which are interleaved at the feature level. The token axis spans the full sequence, including text, special, and audio tokens. The time and band axes explicitly encode the two-dimensional positions of audio tokens, while text tokens use zero band indices and standard sequential time indices. Under band-first flattening, all band tokens within the same frame share the same time index. This design allows the LM to retain temporal and spectral locality during autoregressive decoding.

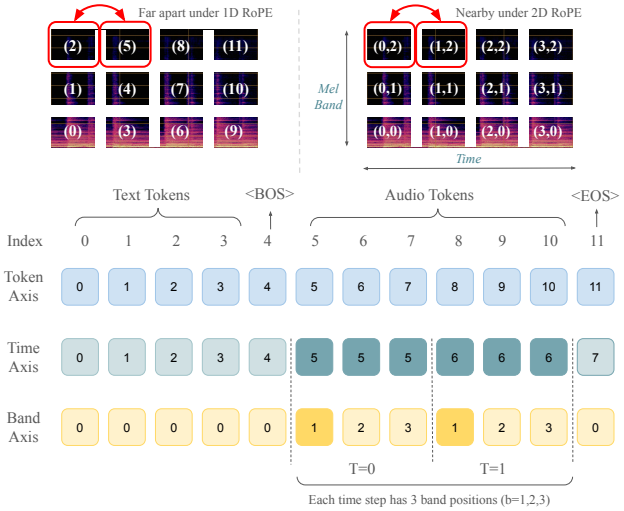

Fig. 3. Illustration of 2D RoPE for flattened audio tokens. It preserves the original time-frequency structure by separately encoding temporal and frequency-band positions. The token axis uses global sequence positions. The time axis follows text-token positions for text tokens and repeats each time-step index across all band tokens. The band axis is set to zero for text tokens and ranges from 1 to B for audio tokens within each time step.

D. Conditioning

We encode the text caption using a pretrained T5 encoder [21] and prepend the resulting embeddings to the audio token sequence. Since captions describe full tracks while training is performed on shorter randomly sampled segments, we additionally encode the segment start time and the total track duration as numerical conditions. These conditions help the model distinguish, for example, an opening segment from a middle segment under the same global caption.

For classifier-free guidance (CFG), following MusicGen [6], we randomly replace the conditioning prefix with a near-null embedding during training. At inference time, we combine the conditional and unconditional logits as

$$\ell_{\text{cfg}} = \ell_{\text{uncond}} + w(\ell_{\text{cond}} - \ell_{\text{uncond}}), \tag{2}$$

where w denotes the guidance scale.

IV. Experiments

In this section, we describe the training details and evaluate the impact of our design choices on reconstruction quality and autoregressive music generation.

A. Datasets

For tokenizer training, we use a mixture of music and general-audio datasets, including FMA [22], Freesound [23], MTG-Jamendo [24], and the MUSDB training set [25]. For language-model training, we use MTG-Jamendo with Qwen2-generated captions from the ICME 2026 Grand Challenge [26]. Since we focus on instrumental music generation, we apply Mel-Band RoFormer [27] for vocal removal and train on the resulting instrumental tracks.

For reconstruction evaluation, we randomly sample 1,000 segments from the MUSDB test set and report Mel and

TABLE I
ABLATION STUDY OF MS-PATCHGAN AND EMA CODEBOOK UPDATES.

Model	Mel ↓	STFT ↓
Baseline w/ PatchGAN	0.837	**1.751**
w/ MS-PatchGAN	**0.749**	1.794
w/ codebook loss	0.763	1.618
w/ EMA	**0.642**	**1.544**

TABLE II
RECONSTRUCTION COMPARISON OF BANDTOK AGAINST BASELINE AUDIO TOKENIZERS.

Model	Bitrate	Mel ↓	STFT ↓
EnCodec-32k	2.2 kbps	1.228	2.300
EnCodec-48k	3.0 kbps	0.942	1.792
EnCodec-48k	6.0 kbps	0.832	1.696
DAC	2.6 kbps†	0.809	1.646
MelCap	2.2 kbps	0.730	1.653
BandTok-1D‡	2.2 kbps	0.690	1.613
BandTok	2.2 kbps	**0.642**	**1.544**

† DAC does not provide an official 2.6 kbps checkpoint; we use the first three quantizer layers from the 8 kbps model to obtain a comparable bitrate. ‡ BandTok-1D denotes the RVQ variant of BandTok.

STFT distances. For generation evaluation, we use the official 100 contest prompts and report $\mathrm{FAD}_{\mathrm{CLAP}}$, $\mathrm{FAD}_{\mathrm{OpenL3}}$, and CLAP score. FAD is computed using CLAP [28] and OpenL3 [29] embeddings with SongDescriber [30] as the reference dataset. We further evaluate on a 586-sample no-singing subset from SongDescriber, following Stable Audio Open [31], and report AudioBox [32] metrics for subjective music-quality assessment, including CE, CU, PC, and PQ, corresponding to Content Enjoyment, Content Usefulness, Production Complexity, and Production Quality, respectively.

B. Reconstruction Improvements

The tokenizer is trained on 8 H800 GPUs for 24 hours with a batch size of 1024 and a segment length of 65,024 samples. We use the Adam optimizer with a learning rate of 2×10^{-4}, $\beta_1 = 0.8$, and $\beta_2 = 0.99$. To stabilize training, we adopt an inverse learning-rate schedule with power 0.5, $\texttt{inv\_gamma} = 200{,}000$, and a warm-up factor of 0.999.

We ablate two reconstruction design choices for BandTok. The multi-scale Mel PatchGAN discriminator applies scale-specific discriminators to spectrograms at different resolutions and improves reconstruction over the standard PatchGAN baseline, as shown in Table I. We also replace the conventional codebook loss with EMA codebook updates while retaining the commitment loss, which stabilizes updates for the single 8192-entry codebook and further improves reconstruction quality. Overall, BandTok achieves better reconstruction than waveform-domain tokenizers, as shown in Table II.

C. Autoregressive Generation

We next evaluate the effect of BandTok on autoregressive language modeling. We focus on two questions: whether two-dimensional Mel-band tokens improve LM modelability, and whether 2D RoPE further improves modeling over flattened time-frequency token sequences.

To isolate the effect of token geometry, we compare BandTok with a variant denoted BandTok-1D, which uses the same model architecture but replaces the vertical Mel-band axis with residual hierarchical codebook layers. For both tokenizations, we train a 315M-parameter language model on 8 H800 GPUs for 19 hours, using a batch size of 128 and 10-second training segments. We use AdamW with a learning rate of 5×10^{-5}, $\beta_1 = 0.9$, and $\beta_2 = 0.95$. We adopt an inverse learning-rate schedule with $\texttt{inv\_gamma} = 1{,}000{,}000$, power 0.5, and a warm-up factor of 0.999.

As shown in Table V, BandTok improves CLAP and $\mathrm{FAD}_{\mathrm{CLAP}}$ over BandTok-1D, indicating that Mel-band tokens are more LM-friendly than residual codebook tokens. We also compare 1D and 2D RoPE for flattened time-frequency sequences. By explicitly encoding temporal and frequency-band positions, 2D RoPE helps the LM preserve the underlying 2D structure and further improves generation quality.

D. Segment-Time Conditioning

Because captions describe full tracks while training uses randomly cropped segments, we add segment-time conditioning, which encodes the segment start time and total track duration following prior work on long-form audio generation [33]. As shown in Table VI, this conditioning improves FAD and CLAP for the 315M model but slightly degrades FAD for the 1.5B model. We hypothesize that larger models rely more strongly on conditions and are therefore more sensitive to mismatches between fixed segment-time settings and varying track structures. Accordingly, we submit models with and without segment-time conditioning as separate variants.

Classifier-free guidance further improves generation quality. Increasing the CFG scale from 1.0 to 2.0 reduces $\mathrm{FAD}_{\mathrm{CLAP}}$ from 0.700 to 0.560 and improves CLAP from 0.148 to 0.186.

E. Token Decoupling Analysis

We analyze whether band-wise tokenization yields a more statistically decoupled token organization. We use normalized mutual information (NMI) as a proxy for pairwise token dependence,

$$\mathrm{NMI}(Z_i, Z_j) = \frac{I(Z_i; Z_j)}{\sqrt{H(Z_i)H(Z_j)}},$$

where lower off-diagonal values indicate weaker statistical coupling across token axes.

We further evaluate autoregressive predictability using a 315M LM under a flattened token modeling scheme. BandTok-1D tokens are flattened along the residual-layer axis, whereas BandTok tokens are flattened along the frequency-band axis. We compute teacher-forced perplexity (PPL) and normalize the per-layer or per-band PPL values to $[0, 1]$.

As shown in Figure 1, residual tokens exhibit stronger coupling and increasing prediction difficulty in later layers. In contrast, although band-wise tokens show a local PPL peak in high-frequency bands, likely due to sparse high-frequency

TABLE III
COMPARISON OF GENERATION PERFORMANCE ACROSS DIFFERENT STAGE-I TOKENIZERS AND STAGE-II GENERATORS ON THE ICME CONTEST TEST SET.

Stage II	Stage I	Params	Train Data (hours)	$\mathbf{FAD_{OpenL3}} \downarrow$	$\mathbf{FAD_{CLAP}} \downarrow$	CLAP ↑
Stable Audio Open	Stable Audio Open VAE	1.1B	7.3k	–	0.574	0.321
MusicGen-small	EnCodec-32k	300M	20k	–	0.574	0.370
MusicGen-medium	EnCodec-32k	1.5B	20k	–	0.548	0.353
MusicGen-large	EnCodec-32k	3.3B	20k	–	0.553	**0.379**
Ours	EnCodec-32k	315M	0.46k	221.327	0.739	0.199
	EnCodec-48k	315M	0.46k	266.994	0.898	0.138
	BandTok	315M	0.46k	163.804	**0.482**	0.163
	BandTok	1.5B	0.46k	**140.006**	0.500	0.171

TABLE IV
GENERATION COMPARISON WITH BASELINE MODELS ON THE SONG DESCRIBER DATASET.

Stage II	Params	CE ↑	CU ↑	PC ↑	PQ ↑
Stable Audio Open	1.1B	6.725	7.634	4.342	7.669
MusicGen-large	3.3B	6.785	7.626	**4.893**	7.498
Ours	315M	6.808	7.627	4.277	7.705
	1.5B	**7.244**	**7.858**	4.040	**7.846**

TABLE V
ABLATION STUDY OF TOKEN GEOMETRY AND POSITIONAL ENCODING FOR AUTOREGRESSIVE GENERATION.

Model	RoPE	$\mathbf{FAD_{CLAP}} \downarrow$	CLAP ↑
BandTok-1D	1D	1.166	0.117
BandTok	1D	0.645	0.193
BandTok	2D	**0.595**	**0.214**

content, they achieve lower inter-token NMI and a more balanced PPL profile across most bands. These results suggest that band-wise tokenization reduces the burden of modeling a residual hierarchy during autoregressive decoding. Both NMI and PPL analyses are conducted on the SongDescriber dataset.

F. Comparison with EnCodec

We further compare BandTok with EnCodec-32k, the waveform tokenizer used in MusicGen. EnCodec-32k represents 32 kHz audio using four 2048-entry codebooks at 50 Hz, yielding 200 tokens per second, comparable to BandTok. Under the same downstream LM architecture, BandTok achieves better generation performance, as shown in Table III.

To examine the potential effect of tokenizer pretraining data, we additionally compare against EnCodec-48k, whose reported training set includes MTG-Jamendo, while EnCodec-32k does not publicly disclose its tokenizer training data. EnCodec-48k represents 48 kHz audio using two 1024-entry codebooks at 150 Hz, yielding 300 tokens per second. As shown in Table III, EnCodec-48k performs worse than EnCodec-32k, likely because its higher token rate and larger token space increase the downstream modeling burden.

TABLE VI
ABLATION STUDY OF CFG AND SEGMENT-TIME CONDITIONING (SEG-TIME COND) FOR DIFFERENT LM SCALES.

Params	Setting	$\mathbf{FAD_{CLAP}} \downarrow$	CLAP ↑
315M	CFG $= 1.0$	0.700	0.148
315M	CFG $= 2.0$	0.560	0.186
315M	+ seg-time cond	**0.509**	**0.206**
1.5B	CFG $= 2.0$	**0.480**	0.217
1.5B	+ seg-time cond	0.486	**0.237**

G. Scaling the Language Model

We study LM scaling by comparing 315M and 1.5B models trained on the same data. As shown in Table III, on the primary evaluation set, increasing model size does not consistently improve $\mathrm{FAD_{CLAP}}$, suggesting that larger LMs may require more training data and more diverse captions. Nevertheless, the 1.5B model improves instrumental timbre recognizability, as reflected by better $\mathrm{FAD_{OpenL3}}$ and CLAP score.

To further assess generation quality, we evaluate on a larger SongDescriber subset. As shown in Table IV, our model achieves the best AudioBox scores among the baselines on this larger set, demonstrating strong generation quality under an academic-scale training setup. While CLAP score remain limited, likely due to the prefix-based text-conditioning strategy, these results highlight the potential of modeling music with two-dimensional time-frequency tokens.

V. CONCLUSION

We presented *BandTok*, a generation-oriented 2D Mel-spectrogram tokenizer for autoregressive music generation. By replacing residual codebook layers with physically meaningful Mel-frequency band tokens, BandTok provides an LM-friendly time-frequency token geometry while maintaining high reconstruction fidelity. With 2D RoPE, BandTok preserves temporal and spectral structure during decoding and improves generation quality over residual-codebook tokenizer baselines. Future work will improve text following through better condition control and caption augmentation, and extend this paradigm to broader audio generation tasks.

REFERENCES

[1] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi, "Soundstream: An end-to-end neural audio codec," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 495–507, 2021.

[2] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi, "High fidelity neural audio compression," *arXiv preprint arXiv:2210.13438*, 2022.

[3] Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar, "High-fidelity audio compression with improved rvqgan," *Advances in Neural Information Processing Systems*, vol. 36, pp. 27980–27993, 2023.

[4] 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*, vol. 31, pp. 2523–2533, 2023.

[5] 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.

[6] 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*, vol. 36, pp. 47704–47720, 2023.

[7] Dongchao Yang, Jinchuan Tian, Xu Tan, Rongjie Huang, Songxiang Liu, Xuankai Chang, Jiatong Shi, Sheng Zhao, Jiang Bian, Xixin Wu, et al., "Uniaudio: An audio foundation model toward universal audio generation," *arXiv preprint arXiv:2310.00704*, 2023.

[8] Jean-Marie Lemercier, Simon Rouard, Jade Copet, Yossi Adi, and Alexandre Défossez, "An independence-promoting loss for music generation with language models," *arXiv preprint arXiv:2406.02315*, 2024.

[9] Jingyi Li, Zhiyuan Zhao, Yunfei Liu, Lijian Lin, Ye Zhu, Jiahao Wu, Qiuqiang Kong, and Yu Li, "Melcap: A unified single-codebook neural codec for high-fidelity audio compression," 2025.

[10] Zhisheng Zhang, Xiang Li, Yixuan Zhou, Jing Peng, Shengbo Cai, Guoyang Zeng, and Zhiyong Wu, "Unisrcodec: Unified and low-bitrate single codebook codec with sub-band reconstruction," *arXiv preprint arXiv:2601.02776*, 2026.

[11] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[12] Yitian Gong, Kuangwei Chen, Zhaoye Fei, Xiaogui Yang, Ke Chen, Yang Wang, Kexin Huang, Mingshu Chen, Ruixiao Li, Qingyuan Cheng, et al., "Moss-audio-tokenizer: Scaling audio tokenizers for future audio foundation models," *arXiv preprint arXiv:2602.10934*, 2026.

[13] Ryan Langman, Ante Jukić, Kunal Dhawan, Nithin Rao Koluguri, and Jason Li, "Spectral codecs: Improving non-autoregressive speech synthesis with spectrogram-based audio codecs," *arXiv preprint arXiv:2406.05298*, 2024.

[14] Yang Ai, Xiao-Hang Jiang, Ye-Xin Lu, Hui-Peng Du, and Zhen-Hua Ling, "Apcodec: A neural audio codec with parallel amplitude and phase spectrum encoding and decoding," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 3256–3269, 2024.

[15] Tao Feng, Zhiyuan Zhao, Yifan Xie, Yuqi Ye, Xiangyang Luo, Xun Guan, and Yu Li, "Stftcodec: High-fidelity audio compression through time-frequency domain representation," in *2025 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2025, pp. 1–6.

[16] Alfred Haar, *Zur theorie der orthogonalen funktionensysteme*, Georg-August-Universitat, Gottingen., 1909.

[17] Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al., "Cosmos world foundation model platform for physical ai," *arXiv preprint arXiv:2501.03575*, 2025.

[18] Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon, "Bigvgan: A universal neural vocoder with large-scale training," *arXiv preprint arXiv:2206.04658*, 2022.

[19] Justin Johnson, Alexandre Alahi, and Li Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *European Conference on Computer Vision*, 2016, pp. 694–711.

[20] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al., "Qwen3-vl technical report," *arXiv preprint arXiv:2511.21631*, 2025.

[21] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020.

[22] Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson, "Fma: A dataset for music analysis," *arXiv preprint arXiv:1612.01840*, 2016.

[23] Eduardo Fonseca, Jordi Pons, Xavier Favory, Frederic Font, Dmitry Bogdanov, Andres Ferraro, Sergio Oramas, Alastair Porter, and Xavier Serra, "Freesound datasets: A platform for the creation of open audio datasets.," in *ISMIR*, 2017, pp. 486–493.

[24] Dmitry Bogdanov, Minz Won, Philip Tovstogan, Alastair Porter, and Xavier Serra, "The mtg-jamendo dataset for automatic music tagging," in *Machine learning for music discovery workshop, international conference on machine learning (ICML 2019)*. Long Beach, CA, United States, 2019, pp. 1–3.

[25] Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner, "Musdb18-hq-an uncompressed version of musdb18," *(No Title)*, 2019.

[26] Fang-Chih Hsieh, Wei-Jaw Lee, Chun-Ping Wang, Hung-yi Lee, Hao-Wen Dong, and Yi-Hsuan Yang, "Academic text-to-music grand challenge: Datasets, baselines, and evaluation methods," in *International Conference on Multimedia and Expo, Grand Challenge Paper*, 2026.

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

[28] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang, "Clap learning audio concepts from natural language supervision," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[29] Jason Cramer, Ho-Hsiang Wu, Justin Salamon, and Juan Pablo Bello, "Look, listen, and learn more: Design choices for deep audio embeddings," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 3852–3856.

[30] Ilaria Manco, Benno Weck, Seungheon Doh, Minz Won, Yixiao Zhang, Dmitry Bogdanov, Yusong Wu, Ke Chen, Philip Tovstogan, Emmanouil Benetos, et al., "The song describer dataset: a corpus of audio captions for music-and-language evaluation," *arXiv preprint arXiv:2311.10057*, 2023.

[31] 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)*. IEEE, 2025, pp. 1–5.

[32] 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.

[33] Zach Evans, CJ Carr, Josiah Taylor, Scott H Hawley, and Jordi Pons, "Fast timing-conditioned latent audio diffusion," in *Forty-first International Conference on Machine Learning*, 2024.