

STABLE AUDIO OPEN

Zach Evans Julian D. Parker CJ Carr Zack Zukowski Josiah Taylor Jordi Pons

Stability AI

ABSTRACT

Open generative models are vitally important for the community, allowing for fine-tunes and serving as baselines when presenting new models. However, most current text-to-audio models are private and not accessible for artists and researchers to build upon. Here we describe the architecture and training process of a new open-weights text-to-audio model trained with Creative Commons data. Our evaluation shows that the model’s performance is competitive with the state-of-the-art across various metrics. Notably, the reported FD_{open13} results (measuring the realism of the generations) showcase its potential for high-quality stereo sound synthesis at 44.1kHz.

Index Terms— Stable Audio Open, Latent Diffusion, Audio.

1. INTRODUCTION

A significant amount of ongoing research focuses on text-conditioned generative audio models [1, 2, 3, 4, 5, 6]. Yet, many models lack public weights or are only available behind private APIs [1, 5, 6], limiting their usefulness as the foundation for further research and artistic creation. Further, the licenses of the audio used for training public models are often not fully documented. For example, AudioGen [2] or AudioLDM [3, 4] are trained on a mix of public datasets that include AudioSet [7], and we are unaware of any disclosed licenses for AudioSet’s audio. MusicGen [8], on the other hand, is an open model with well-documented training data and licences, trained exclusively on licensed¹ copyrighted data. Finally, current open models are not competitive against the state-of-the-art in terms of quality, coherent generation over long sequences, and inference speed [5, 6]. Given that, our goal is to release a text-conditioned generative model for non-speech audio based on the following:

- Trained only on Creative Commons (CC) licensed audio.
- Publicly available model weights and code, along with the attributions needed for the used data, to facilitate open research.
- State-of-the-art sound quality generation at 44.1kHz stereo.

While this data choice may limit our model’s capabilities (specially for text-to-music as noted in section 6.1) it facilitates transparent data practices at the time of releasing the model. This manuscript describes how these goals were achieved, including the description of our architecture (section 2), how we handle training data and ensure that such audio is not copyrighted (section 3), and how our model was trained (section 4). In section 5 we evaluate the resulting model and its parts (the generative model and the autoencoder) by a number of standard metrics and study whether it exhibits memorization of the training data. We also show that our model can run on consumer-grade GPUs. Our aim is to further improve current best practices for open model releases, with an emphasis on evaluation, data transparency, and accessibility for artists and scholars.

¹Documented in: https://github.com/facebookresearch/audiocraft/blob/main/model_cards/MUSICGEN_MODEL_CARD.md

2. ARCHITECTURE

Our latent diffusion model generates variable-length (up to 47s) stereo audio at 44.1kHz from text prompts. It consists of 3 parts: an autoencoder (156M parameters) that compresses waveforms into a manageable sequence length, a T5-based text embedding [9] for text conditioning (109M parameters), and a transformer-based diffusion model (DiT of 1057M parameters) that operates in the latent space of the autoencoder. Our model is a variant of Stable Audio 2.0 [6] that is trained on CC data. Its architecture is similar except that it uses T5 [9] text conditioning instead of CLAP [10]. The model’s exact parameterization and its weights are available online.²

2.1. Autoencoder

The (variational) autoencoder operates on raw waveforms. The encoder processes such waveforms with 5 convolutional blocks, each performing downsampling and channel expansion via strided convolutions. Before each downsampling block, we employ a series of ResNet-like layers using dilated convolutions and Snake [11] activation functions for further processing. The bottleneck of the autoencoder is parameterized as a variational autoencoder with a latent size of 64. The decoder is an inversion of the encoder structure, employing transposed strided convolutions for upsampling, with channel contraction before each upsampling block. All convolutions are weight-normalised and the output of the decoder does not include a $\tanh()$ as we found it introduced harmonic distortion.

2.2. Diffusion-transformer (DiT)

Our generative model is a diffusion-transformer (DiT) [6, 12, 13] that follows a standard structure with stacked blocks consisting of serially connected attention layers and gated multi-layer perceptrons (MLPs), with skip connections around each. We employ bias-less layer normalization at the input to both the attention layer and the MLP. The key and query inputs to the attention layer have rotary positional embeddings [14] applied to only half of the embeddings. Each transformer block also contains a cross-attention layer to incorporate conditioning. Linear mappings are used at the input and output of the transformer to translate from the autoencoder latent dimension to the embedding dimension of the transformer. Efficient block-wise attention [15] and gradient checkpointing [16] are employed to reduce the computational and memory requirements.

The DiT is conditioned by 3 signals: *text* enabling language control, *timing* enabling variable-length generation, and *timestep* signaling the current diffusion timestep. Text conditioning is provided by the pretrained T5-base encoder [9]. Timestep conditioning [5, 17] goes through sinusoidal embeddings [18]. Conditioning is introduced via cross-attention or via prepending conditioning signals to the input. Cross-attention includes timing and text conditioning. Prepend conditioning includes timing and timestep conditioning.

²<https://huggingface.co/stabilityai/stable-audio-open-1.0/>

2.3. Variable-length audio generation

As natural audio can be of various lengths, our model supports variable-length audio generation within a specified window (e.g., 47s) by relying on the timing condition to fill the signal up to the specified length. The model is trained to fill the rest with silence. To present variable-length audios (shorter than the window length) to the end-users, one can easily trim the appended silence. We adopt this strategy, as it has previously shown its effectiveness [5, 6].

3. TRAINING DATA

Our dataset consists of CC recordings from Freesound and the Free Music Archive (FMA). We conducted an analysis to ensure no copyrighted content was in our training data. To that end, we first identified music recordings in Freesound using the PANNs [19] tagger. The identified music activated music-related tags for at least 30 sec (threshold of 0.15). This threshold was set with FMA music examples and ensuring no false negatives were present. The identified music was sent to a trusted content detection company to ensure the absence of copyrighted music. The flagged copyrighted music was subsequently removed from our training set. Most of the removed content was field recordings in which copyrighted music was in the background. Following this procedure, we are left with 266,324 CC0, 194,840 CC-BY, and 11,454 CC Sampling+ audio recordings.

We also conducted an analysis to ensure no copyrighted content was present in FMA’s subset. In this case, the procedure was different because the FMA subset consists of only music. We did a metadata search against a large database of copyrighted music and flagged any potential match to be reviewed by humans. After this process, we ended up with 8,967 CC-BY and 4,907 CC0 tracks.

This led to a dataset with 486,492 recordings (7,300 h), where 472,618 (6,330 h) are from Freesound and 13,874 (970 h) from FMA, all licensed under CC-0, CC-BY, or CC-Sampling+. The most common tags in those Freesound recordings are *single-note*, *synthesizer*, *field-recording*, *drum*, *loop*, *ambient* and in FMA are *instrumental*, *electronic*, *experimental*, *soundtrack*, *ambient*. This data is used to train most of our system (autoencoder and DiT, not the publicly available T5-base [9] that was pretrained) from scratch.

3.1. Autoencoder training data

We gathered 5 sec chunks of diverse, high fidelity audio. First, we gathered up to $\times 3$ (random) chunks for each Freesound audio, to ensure diversity and avoid oversampling long recordings. Then, we gathered additional (random) chunks from a high fidelity subset, including all FMA tracks and a subset of stereo and full-band Freesound audio (55,314). Such high fidelity Freesound audio were recorded at 48kHz and verified to contain energy in high frequencies. FMA tracks were variable-bitrate MP3s encoded at 44.1kHz. Note that high fidelity Freesound recordings were sampled twice.

3.2. Prompts preparation for training the DiT

Training audio is paired with text metadata. Freesound examples include natural language descriptions as well as the title of the recording and tags. FMA music examples include metadata like year, genres, album, title, and artist. We generate text prompts from the metadata by concatenating a random subset of the metadata as a string. This allows for specific properties to be specified during inference, while not requiring these properties to be present at all times. For metadata-types with a list of values, like tags or genres, we shuffle

the list. As a result, we perform a variety of random transformations to the resulting string, shuffling the order and also transforming between upper and lower case. For half of the FMA prompts, we include the metadata-type (e.g., artist or album) and join them with a comma (e.g., “*year: 2021, artist: dadabots, album: can’t play instruments, title: pizza hangover*”). For the other half, we do not include the metadata-type and join them with a comma (e.g., “*dadabots, can’t play instruments, pizza hangover, 2021*”).

4. EXPERIMENTAL SETUP

All models are trained with AdamW, with weight decay of 0.001 and a learning rate scheduler including exponential ramp-up and decay. The exact training hyperparameters we used are detailed online.²

4.1. Autoencoder training

We train the variational autoencoder using a variety of objectives. First, there is a reconstruction loss, based on a perceptually weighted multi-resolution STFT [23] that deals with stereo signals via the mid-side (M/S) representation of the stereo audio, as well as the left and right (L/R) channels. The L/R component is down-weighted by 0.5 compared to the M/S component, and exists to mitigate ambiguity around left-right placement. Second, we employ an adversarial loss term, utilizing 5 convolutional discriminators as in Encodec [22]. Third, the KL divergence loss term is down-weighted by 1×10^{-4} .

It trained for 183 h on $\times 32$ A100s with a batch size of 4. At this point the encoder was frozen, and the decoder was trained for another 273 h on $\times 32$ A100s with a batch size of 8. Each batch is made of ≈ 1.5 sec chunks (65,536 samples at 44.1kHz). The autoencoder itself was trained with a base learning rate of 1.5×10^{-4} , and the discriminators with a learning rate of 3×10^{-4} .

4.2. DiT training and inference

The DiT is trained to predict a noise increment from noised ground-truth latents, following the v-objective [24] approach. During inference, we use the DPM-Solver++ [25] for 100 steps with classifier-free guidance (scale of 0.7). The DiT is trained for 338 h on $\times 64$ A100s with a batch size of 4 and a base learning rate of 5×10^{-5} . Each batch contains latent sequences of length 1024 (≈ 47 sec).

5. EVALUATION

5.1. Generative model evaluation

We employ established quality metrics³ that include FD_{openl3} [26], KL_{pass} [27] and $CLAP_{score}$ [10, 28]. A low FD_{openl3} implies that the generated audio is plausible and closely matches the reference [29, 8]. A low KL_{pass} indicates semantic correspondence between the generated and the reference audio [8]. A high $CLAP_{score}$ denotes that the generated audio adheres to the given text prompt [10, 28]. We use two evaluation sets: AudioCaps Dataset [30] for sound generation, and Song Descriptor Dataset [31] for music generation.

Table 1 shows the AudioCaps Dataset results. AudioCaps’ test set contains 979 audio segments from YouTube, each with several captions (881 audios were available and it includes 4,875 captions). We generate an audio per caption, resulting in 4,875 generations. We use previous Stable Audio [5, 6] models as baselines. We also include AudioLDM2 [4] and AudioGen [20] as a reference, since they are the most competitive open models available, but those might not

³<https://github.com/Stability-AI/stable-audio-metrics>

| | channels/sr | output length | FD _{openl3} ↓ | KL _{passt} ↓ | CLAP _{score} ↑ |
|----------------------|-------------|---------------|------------------------|-----------------------|-------------------------|
| AudioLDM2-48kHz [4] | 1/48kHz | 10 sec | 101.11 | 2.04 | 0.37 |
| AudioLDM2-large [4] | 1/16kHz | 10 sec | 170.31 | 1.57 | 0.41 |
| AudioGen-medium [20] | 1/16kHz | 10 sec | 186.53 | 1.42 | 0.45 |
| Stable Audio 1.0 [5] | 2/44.1kHz | 95 sec † | 103.66 | 2.89 | 0.24 |
| Stable Audio 2.0 [6] | 2/44.1kHz | 190 sec † | 116.14 | 2.67 | 0.24 |
| Stable Audio 2.0 [6] | 2/44.1kHz | 285 sec † | 110.62 | 2.70 | 0.23 |
| Stable Audio Open | 2/44.1kHz | 47 sec † | 78.24 | 2.14 | 0.29 |

Table 1. *AudioCaps Dataset (with sounds, not music)*. Stable Audio Open outperforms comparable baselines, showcasing its potential in synthesizing sounds and field recordings. Since AudioCaps is a subset of AudioSet, our results might not be comparable to AudioLDM2/AudioGen as those are trained with AudioSet. † Stable Audio models are trained to generate longer outputs, but during inference can generate variable-length outputs relying on the timing conditioning. We trim the generated audio to 10 sec (discarding the end silent part) for a fair comparison.

| | channels/sr | output length | FD _{openl3} ↓ | KL _{passt} ↓ | CLAP _{score} ↑ |
|---------------------------|-------------|---------------|------------------------|-----------------------|-------------------------|
| MusicGen-large-stereo [8] | 2/32kHz | 47 | 190.47 | 0.52 | 0.31 |
| Stable Audio 1.0 [5] | 2/44.1kHz | 95 sec † | 142.50 | 0.40 | 0.38 |
| Stable Audio 2.0 [6] | 2/44.1kHz | 190 sec † | 71.25 | 0.37 | 0.42 |
| Stable Audio 2.0 [6] | 2/44.1kHz | 285 sec † | 81.05 | 0.39 | 0.42 |
| Stable Audio Open | 2/44.1kHz | 47 | 96.51 | 0.55 | 0.41 |

Table 2. *Song Describer Dataset (instrumental music, not sounds)*. Stable Audio Open is worse than Stable Audio at generating music, but slightly better than MusicGen (best open model). † Stable Audio models are trained to generate longer outputs, but during inference can generate variable-length outputs relying on the timing conditioning. We trim the generated audio to 47 sec (discarding the end silent part) for a fair comparison.

| | sampling rate | STFT distance ↓ | MEL distance ↓ | SI-SDR ↑ | latent rate | latent (channels) |
|----------------------|---------------|-----------------|----------------|-------------|-------------|-------------------|
| DAC [21] | 44.1kHz | 1.04 | 0.63 | 5.05 | 86Hz | discrete |
| AudioGen [20] | 48kHz | 1.10 | 0.62 | 4.60 | 50Hz | discrete |
| Codec [22, 8] | 32kHz | 1.82 | 1.12 | 5.33 | 50Hz | discrete |
| AudioGen [20] | 48kHz | 1.22 | 0.73 | 2.04 | 100Hz | continuous (32) |
| Stable Audio 1.0 [5] | 44.1kHz | 1.32 | 0.86 | 1.81 | 43Hz | continuous (64) |
| Stable Audio 2.0 [6] | 44.1kHz | 1.21 | 0.85 | 0.33 | 21.5Hz | continuous (64) |
| Ours | 44.1kHz | 1.25 | 0.86 | -0.93 | 21.5Hz | continuous (64) |

Table 3. *Autoencoder reconstructions: AudioCaps Dataset (sounds)*. Different autoencoders use various sampling rates but evaluations are conducted at 44.1kHz for a fair comparison. Also note that (i) continuous latents are not comparable to discrete ones, and (ii) different latent rates are not strictly comparable. Sorted by latent rate.

| | sampling rate | STFT distance ↓ | MEL distance ↓ | SI-SDR ↑ | latent rate | latent (channels) |
|----------------------|---------------|-----------------|----------------|-------------|-------------|-------------------|
| DAC [21] | 44.1kHz | 0.96 | 0.53 | 11.30 | 86Hz | discrete |
| AudioGen [20] | 48kHz | 1.16 | 0.66 | 9.64 | 50Hz | discrete |
| Codec [22, 8] | 32kHz | 1.70 | 1.09 | 5.75 | 50Hz | discrete |
| AudioGen [20] | 48kHz | 1.10 | 0.65 | 9.21 | 100Hz | continuous (32) |
| Stable Audio 1.0 [5] | 44.1kHz | 1.20 | 0.67 | 9.12 | 43Hz | continuous (64) |
| Stable Audio 2.0 [6] | 44.1kHz | 1.21 | 0.72 | 7.65 | 21.5Hz | continuous (64) |
| Ours | 44.1kHz | 1.29 | 0.77 | 6.28 | 21.5Hz | continuous (64) |

Table 4. *Autoencoder reconstructions: Song Describer Dataset (instrumental music)*. Different autoencoders use various sampling rates but evaluations are conducted at 44.1kHz for a fair comparison. Also note that (i) continuous latents are not comparable to discrete ones, and (ii) different latent rates are not strictly comparable. Sorted by latent rate.

be strictly comparable since they were trained on AudioSet (and AudioCaps is a subset of AudioSet). Stable Audio Open outperforms comparable baselines, particularly on FD_{open13} , highlighting its potential to generate realistic sounds and field recordings.

Table 2 shows the Song Describer Dataset results. We select this benchmark because others contain bad quality music and shorter snippets [1]. As vocal generation is not our focus and baselines are not trained for that, we run a fair evaluation using a subset of prompts describing instrumental music (no-singing) [5, 6] with 586 captions for 446 music tracks. We generate an audio per caption, which results in 586 generations. We set previous Stable Audio [5, 6] models and MusicGen-large-stereo [8] as baselines. The latter is the most competitive open model for stereo music generation [5]. Our results suggest that Stable Audio Open is worse than Stable Audio at generating music but slightly better than MusicGen (best open model).

5.2. Autoencoder evaluation

Note that the autoencoder can also be used alone and is implicitly released with the model. For this reason we also evaluate its audio reconstruction quality (Tables 3 and 4) by comparing ground-truth and reconstructed audio via a set of established audio quality metrics [21, 22]: STFT distance, MEL distance and SI-SDR (as in auraloss library [23], with its default parameters). The reconstructed audio is obtained by encoding-decoding the ground-truth audio from the AudioCaps Dataset (881 recordings) and Song Describer Dataset (no-singing subset with 446 tracks). We compare our autoencoder against a number of neural audio codecs including Encodec [22], DAC [21] and AudioGen [20]. We select the Encodec 32kHz variant because the MusicGen-large-stereo baseline relies on it, and DAC 44.1kHz because its alternatives operate at 24kHz and 16kHz. Further, as our autoencoder relies on a continuous latent, we also compare with AudioGen, a state-of-the-art autoencoder with both continuous and discrete latent options. All our baselines are stereo, except DAC and Encodec. In those cases, we independently project left and right channels and reconstruct from those. Results show that lower latent rates generally yield worse reconstructions. Considering this, note that our model is only comparable to Stable Audio 2.0 (being continuous, with the same latent rate) and shows nearly similar performance despite being trained only with CC data.

5.3. Memorization (exact copies) analysis

Recent works [6, 32, 33] examined the potential of generative models to memorize training data, especially for repeated elements in the training set. Further, MusicLM [1] conducted a memorization analysis to address concerns on the potential misappropriation of creative content. Adhering to principles of responsible model development, we also run a comprehensive study on memorization [1, 6, 32, 33].

In light of the possible risk of memorizing repeated audio within the training set, we start by studying if our dataset contains repeated data. We embed all our training data using the LAION-CLAP [10] audio encoder to select audios that are close in this space based on a manually set threshold. The threshold is set such that the selected audio correspond to exact replicas. With this process, we identify 3,693 Freesound and 856 FMA repeated recordings.

Our methodology is based on comparing our model’s generations against the training set in LAION-CLAP space. We then select the top-50 generations that are closest to the training data (the memorization candidates) and listen. We listened to memorization candidates generated with prompts from the identified repeated data in our training set, and did not find memorization. We also listened to mem-

orization candidates from 11,000 random prompts from the training set, and did not find memorization. We even listened to memorization candidates from outstanding generations, and did not find memorization. The most interesting memorization candidates, together with their closest training data, are online for listening⁴. Those include similar generations of well defined sounds like “*storm*” or “*1000Hz*”, but we did not find memorization beyond that.

Also note that our model only generates 47s audio and cannot memorize longer audio. Further, it cannot produce intelligible speech or singing, making it difficult to memorize such examples.

Finally, note that the primary objective of generative modeling is to create new content based on the training data. Simply reproducing the training data indicates poor performance and is not interesting.

5.4. Inference speed on various hardware

Our model runs 8 inference steps per second (steps/sec) on an RTX-3090 (24GB VRAM), 11 steps/sec on an RTX-A6000 (48GB), and 20 steps/sec on an H100 (80GB). Appendices B and C include more details on VRAM usage and on which hardware it can be finetuned.

6. CONCLUSIONS

This article documents the release of Stable Audio Open with a particular focus on evaluation and data transparency. Our results show its potential for synthesizing high-quality stereo sounds at 44.1kHz. Further, we also release a continuous autoencoder that operates at a low latent rate (21.5Hz) that can work for both music and audio. Our model is accessible to everyone² and can run on consumer-grade GPUs, making it appealing for both academic and artistic use cases.

6.1. Limitations

On audio generation. Our model finds challenging to generate prompts with connectors⁵ and cannot generate intelligible speech⁶. It sometimes omits one (or more) of the sounds present in prompts with connectors, e.g., “*A man speaking as a crowd of people laugh and applaud*” where the generated sound includes a man speaking with mild background noise, but with no laughter or applause. Also, speech generations are not intelligible because our model is not “spoken-word” conditioned. Table 5 shows that $CLAP_{score}$ improves when removing connectors- and speech-related prompts.

| | $CLAP_{score}$ (\uparrow) | # files |
|--------------------------|-------------------------------|---------|
| AudioCaps | 0.29 | 4,875 |
| – no speech prompts | 0.31 | 2,765 |
| – no connectors prompts | 0.32 | 711 |
| – no speech & connectors | 0.34 | 587 |

Table 5. $CLAP_{score}$ depending on the used prompts.

On music generation. Note that most commercial music is copyrighted. Hence, our model was trained with limited high-quality music since we focused on CC training data. As a result, it is not competitive against state-of-the-art music models (Tables 2 and 4).

On prompting. Due to the above limitations, prompt engineering may be required for best results. Further, it was mainly trained with English text and is not expected to perform well in other languages.

⁴<https://stability-ai.github.io/stable-audio-open-demo/>

⁵Including: *and, followed, while, as, then, with, later, before, after.*

⁶Including: *speech, male, female, woman, man, speaking, speaks.*

7. REFERENCES

- [1] Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank, “MusicLM: Generating music from text,” *arXiv*, 2023.
- [2] Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi, “Audiogen: Textually guided audio generation,” *arXiv*, 2022.
- [3] Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley, “AudioLDM: Text-to-audio generation with latent diffusion models,” *arXiv*, 2023.
- [4] Haohe Liu, Qiao Tian, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D. Plumbley, “AudioLDM 2: Learning holistic audio generation with self-supervised pretraining,” *arXiv*, 2023.
- [5] Zach Evans, CJ Carr, Josiah Taylor, Scott H. Hawley, and Jordi Pons, “Fast timing-conditioned latent audio diffusion,” *ICML*, 2024.
- [6] Zach Evans, Julian D Parker, CJ Carr, Zack Zukowski, Josiah Taylor, and Jordi Pons, “Long-form music generation with latent diffusion,” *arXiv*, 2024.
- [7] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter, “AudioSet: An ontology and human-labeled dataset for audio events,” in *ICASSP*, 2017.
- [8] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez, “Simple and controllable music generation,” *arXiv*, 2023.
- [9] 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,” *The Journal of Machine Learning Research*, 2020.
- [10] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov, “Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation,” *ICASSP*, 2023.
- [11] Liu Ziyin, Tilman Hartwig, and Masahito Ueda, “Neural networks fail to learn periodic functions and how to fix it,” *arXiv*, 2020.
- [12] William Peebles and Saining Xie, “Scalable diffusion models with transformers,” in *ICCV*, 2023.
- [13] Mark Levy, Bruno Di Giorgi, Floris Weers, Angelos Katharopoulos, and Tom Nickson, “Controllable music production with diffusion models and guidance gradients,” *arXiv*, 2023.
- [14] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu, “RoFormer: Enhanced transformer with rotary position embedding,” *arXiv*, 2023.
- [15] Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré, “FlashAttention: Fast and memory-efficient exact attention with IO-awareness,” *arXiv*, 2022.
- [16] Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin, “Training deep nets with sublinear memory cost,” *arXiv*, 2016.
- [17] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever, “Jukebox: A generative model for music,” *arXiv*, 2020.
- [18] Jonathan Ho, Ajay Jain, and Pieter Abbeel, “Denoising diffusion probabilistic models,” *arXiv*, 2020.
- [19] Qiuqiang Kong, Yin Cao, Turab Iqbal, Yuxuan Wang, Wenwu Wang, and Mark D Plumbley, “PANNs: Large-scale pre-trained audio neural networks for audio pattern recognition,” *IEEE/ACM TASLP*, vol. 28, pp. 2880–2894, 2020.
- [20] AudiogenAI, “Audiogenai/agg: Audiogen codec,” .
- [21] Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar, “High-fidelity audio compression with improved rvqgan,” *NeurIPS*, 2024.
- [22] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi, “High fidelity neural audio compression,” *arXiv*, 2022.
- [23] Christian J. Steinmetz and Joshua D. Reiss, “auraloss: Audio focused loss functions in PyTorch,” in *DMRN+15*, 2020.
- [24] Tim Salimans and Jonathan Ho, “Progressive distillation for fast sampling of diffusion models,” *arXiv*, 2022.
- [25] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu, “DPM-solver++: Fast solver for guided sampling of diffusion probabilistic models,” *arXiv*, 2022.
- [26] Aurora Linh Cramer, Ho-Hsiang Wu, Justin Salamon, and Juan Pablo Bello, “Look, listen, and learn more: Design choices for deep audio embeddings,” *ICASSP*, 2019.
- [27] Khaled Koutini, Jan Schlüter, Hamid Eghbal-zadeh, and Gerhard Widmer, “Efficient training of audio transformers with patchout,” *INTERSPEECH*, 2022.
- [28] Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao, “Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models,” *arXiv*, 2023.
- [29] Kevin Kilgour, Mauricio Zuluaga, Dominik Roblek, and Matthew Sharifi, “Fréchet Audio Distance: A metric for evaluating music enhancement algorithms,” *arXiv*, 2018.
- [30] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim, “AudioCaps: Generating captions for audios in the wild,” *Conf. of the North American Chapter of the Association for Computational Linguistics*, 2019.
- [31] Iaria Manco, Benno Weck, SeungHeon Doh, Minz Won, Yixiao Zhang, Dmitry Bogdanov, Yusong Wu, Ke Chen, Philip Tovstogan, Emmanouil Benetos, Elio Quinton, György Fazekas, and Juhan Nam, “The Song Descriptor Dataset: a corpus of audio captions for music-and-language evaluation,” *arXiv*, 2023.
- [32] Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwal, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace, “Extracting training data from diffusion models,” in *USENIX Security Symposium*, 2023.
- [33] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al., “Scaling rectified flow transformers for high-resolution image synthesis,” *arXiv*, 2024.

| | channels/sr | output length | FD_{open13} ↓ | KL_{passt} ↓ | $CLAP_{score}$ ↑ |
|---------------------------|-------------|---------------|-----------------|----------------|------------------|
| MusicGen-large-stereo [8] | 2/32kHz | 47 | 193.98 | 0.54 | 0.30 |
| Stable Audio 1.0 [5] | 2/44.1kHz | 95 sec † | 139.41 | 0.36 | 0.40 |
| Stable Audio 2.0 [6] | 2/44.1kHz | 190 sec † | 72.17 | 0.35 | 0.44 |
| Stable Audio 2.0 [6] | 2/44.1kHz | 285 sec † | 80.44 | 0.36 | 0.43 |
| Stable Audio Open | 2/44.1kHz | 47 | 99.70 | 0.62 | 0.43 |

Table 6. *Song Describer Dataset (all dataset with music, not sounds)*. Stable Audio Open is worse than Stable Audio at generating music, but slightly better than MusicGen (best open model). † Stable Audio models are trained to generate longer outputs, but during inference can generate variable-length outputs relying on the timing conditioning. We trim the generated audio to 47 sec (discarding the end silent part) for a fair comparison.

| | sampling rate | STFT distance ↓ | MEL distance ↓ | SI-SDR ↑ | latent rate | latent (channels) |
|----------------------|---------------|-----------------|----------------|-------------|-------------|-------------------|
| DAC [21] | 44.1kHz | 0.96 | 0.52 | 10.83 | 86Hz | discrete |
| AudioGen [20] | 48kHz | 1.17 | 0.64 | 9.27 | 50Hz | discrete |
| Codec [22, 8] | 32kHz | 1.82 | 1.12 | 5.33 | 50Hz | discrete |
| AudioGen [20] | 48kHz | 1.10 | 0.64 | 8.82 | 100Hz | continuous (32) |
| Stable Audio 1.0 [5] | 44.1kHz | 1.19 | 0.67 | 8.62 | 43Hz | continuous (64) |
| Stable Audio 2.0 [6] | 44.1kHz | 1.19 | 0.71 | 7.14 | 21.5Hz | continuous (64) |
| Ours | 44.1kHz | 1.32 | 0.78 | 5.78 | 21.5Hz | continuous (64) |

Table 7. *Autoencoder reconstructions: Song Describer Dataset (all dataset)*. Different autoencoders use various sampling rates but evaluations are conducted at 44.1kHz for a fair comparison. Also note that (i) continuous latents are not comparable to discrete ones, and (ii) different latent rates are not strictly comparable. Sorted by latent rate.

A. ADDITIONAL SONG DESCRIBER DATASET RESULTS

Tables 2 and 4 show the results for a subset of the Song Describer Dataset that includes only prompts for instrumental music¹ [5, 6]. As vocal generation is not our focus and baselines were not trained for that either, the main body of the paper reports results on the instrumental subset to ensure a fair evaluation. Yet, for completeness, this appendix also includes results on the entire Song Describer Dataset (Tables 6 and 7) with the same metrics as in Tables 2 and 4. Note that although results vary slightly, the trends remain consistent.

A.1. Generative model evaluation

Table 6 shows the Song Describer Dataset (all dataset) results. The dataset contains 1,106 captions for 706 tracks [31]. We generate an audio per caption, which results in 1,106 generations. Our results suggest that Stable Audio Open is worse than Stable Audio at generating music, but slightly better than MusicGen (best open model).

A.2. Autoencoder evaluation

Table 7 shows the Song Describer Dataset (all dataset) results. We evaluate the reconstructed audio obtained by encoding-decoding the 706 tracks [31] from the dataset. Results show that lower latent rates generally yield worse reconstructions. Considering this, note that our model is only comparable to Stable Audio 2.0 (being continuous, with the same latent rate) and shows slightly worse performance despite being trained only with CC data.

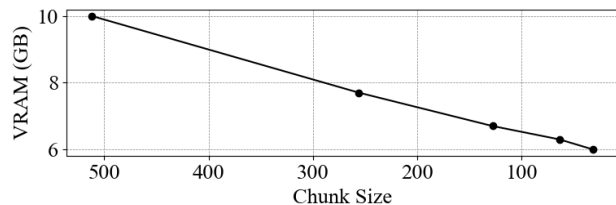
¹Prompts containing any of those words were removed: *speech, speech synthesizer, hubbub, babble, singing, male, man, female, woman, child, kid, synthetic singing, choir, chant, mantra, rapping, humming, groan, grunt, vocal, vocalist, singer, voice, and acapella*.

B. VRAM CONSUMPTION DURING INFERENCE

During diffusion the DiT utilizes 5.9 GB VRAM. During decoding, rendering waveforms from latents, RAM usage increases to 14.5 GB.

B.1. Chunk decoding

Decoding waveforms from latents consumes significantly more VRAM than the DiT. To reduce the memory footprint of the decoder, we explore chunk decoding by splitting the latent sequence into overlapping chunks, decoding them separately, and reassembling them into a final audio. As long as overlaps include the decoder’s receptive field (16 latents on each side), the resulting audio is the same. The VRAM usage after chunking is as follows:



C. FINETUNING ON VARIOUS HARDWARE

@*RoyalCities* changed the window length from 47 to 20 sec, and finetuned it on piano loops using $\times 2$ RTX-A6000 (48GB)². @*Lyraaaa* changed the window length from 47 to 11 sec, and finetuned it on loops/oneshots using $\times 1$ or $\times 2$ RTX-A6000 (48GB)³.

²<https://x.com/RoyalCities/status/1808563794677018694>

³https://www.youtube.com/watch?v=ex4OBD_lrds