Efficient Fine-Tuning of CLAP-Guided Latent Stable Diffusion for High Fidelity Music Generation

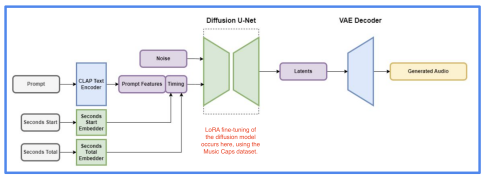
CS 2420: Computing at Scale
Elizabeth Li*†‡, Jamin Liu*†‡, Gabe Mehra*†

*Harvard College
†Harvard John A. Paulson School of Engineering and Applied Sciences
‡Harvard Kenneth C. Griffin Graduate School of Arts and Sciences

***Abstract*—Generative text-to-audio models hold immense potential for enhancing creativity in music-making. However, existing models often lack the specificity required for high-quality music generation. This project addresses this gap by fine-tuning Stability AI's Stable Audio model—a CLAP-guided latent diffusion model—on the Music Caps dataset using Low-Rank Adaptation (LoRA). This approach not only enhances the model's ability to generate musically coherent audio but also significantly reduces computational costs. Our results, evaluated through both objective metrics and subjective listener surveys, demonstrate improvements in musical quality and training efficiency, contributing to the growing body of research in music-focused generative models.**

I. Introduction

Generative music models offer a promising avenue for enhancing the creative process in music-making. Text-to-audio models, in particular, can serve as collaborative tools, providing stylistic ideas or phrases to inspire musicians. While existing open-source text-to-audio models vary in their audio generation capabilities, few are fine-tuned specifically for high-quality music generation. Consequently, general-purpose audio models often fail to produce outputs of high musical integrity, highlighting the need for improved models tailored to music generation.



LoRA-aided fine-tuning on audio diffusion.

Inspired by OpenAI's Contrastive Language-Image Pretraining (CLIP) model [1], which guides stable diffusion for image generation, we propose an analogous approach for audio. Specifically, we leverage Microsoft's Contrastive Language-Audio Pretraining (CLAP) model [2] and Stability AI's Stable Audio framework [3], which integrates CLAP's text encoder with stable diffusion, to create a text-to-audio generation pipeline. However, fine-tuning Stable Audio's diffusion process directly incurs significant computational costs. To address this, our project employs Low-Rank Adaptation (LoRA) [4] to fine-tune the model on Music Caps, a music-specific dataset, with greatly reduced training overhead. Doing so achieves two key objectives: first, it fine-tunes a text-to-audio model specifically for music generation on a dataset where this has not been attempted, and second, it ensures computational efficiency.

II. Related Work

The Stable Audio model is a generative text-to-audio model that combines CLAP for its text-to-audio functionality, U-Net diffusion for audio generation, and a varational autoencoder for the ability to represent audio files in latent space [2] [3]. Model fine-tuning pulls from LoRA, a technique introduced in class that involves training low-rank adaptation matrices to efficiently fine-tune pre-trained models without modifying the original weights [4]. There are also existing models in the literature that have fine-tuned Stable Audio for generation of specific sounds [5], but no open-source fine-tune exists for our dataset. And in terms of evaluation metrics of our generated music, we pull from existing literature that details both objective and subjective evaluation metrics, quantifying the harmonic, rhythmic, and genre-related accuracy of music generated through large language models [6] [7] [8].

III. Approach

We fine-tune CLAP-guided diffusion in the original Stable Audio model for higher quality music generation by using additional audio-text data, and to mitigate the issue of expensive fine-tuning, we apply our *secret weapon LoRA* to fine-tune Stable Audio's diffusion model, which is a U-Net that constructs latent audio representations. In particular, using LoRA freezes the original model and injects LoRA-trainable matrices into the cross-attention layers of the diffusion model, necessitating saving only the weights of the new layers derived from fine-tuning the "residual" of the model instead of all trainable weights. Thus, we are able to reduce computation drastically.

Our approach is novel as no pre-existing literature fine-tunes Stable Audio's diffusion model using our specific dataset (detailed below IV). Our LoRA fine-tuning of this diffusion model using our dataset is therefore novel as well. The novelty of our approach is owed in part to the novelty of the field —

most research on audio diffusion models has been published within the last 6 months [3].

IV. Implementation

We now discuss the tangible implementation of our approach through detailing our tools and walking through our concrete experimental setup.

A. Models, Dataset, and Tools

We utilize the Stable Audio Open 1.0 model [5] for our fine tuning. We select this model as it is the most recent audio diffusion model open-sourced on HuggingFace. Additionally, the model effectively combines a text-to-audio model (CLAP) with stable diffusion (U-Net diffusion) and an encoding-decoding scheme for latent audio representation (variational audio encoder and decoder) in one streamlined model, allowing for a streamlined workflow to fine-tune stable diffusion directly, eliminating the need to independently integrate and manage these components.

Our dataset of choice is the Music Caps dataset from Google Research [9], an openly available collection of audio files and text captions (genre, key, tempo, stylistic description, etc.) from over 5000 diverse copyright free music tracks curated on YouTube. We select this dataset for its novelty, and also for its pre-pairing of .wav files with accompanying text metadata. We access the models and dataset through Hugging Face [5] [9], and to make these files, in addition to our own code and outputs, accessible to our group members, we conducted file management using Google Drive (linked in Section VIII).

To implement LoRA fine-tuning, we use previously existing architectures for fine-tuning stable audio diffusion accessible through GitHub [10]. For our evaluation metric implementation, we use pre-existing GitHub repositories as well [6].

B. Experimental Set-Up

To evaluate the quality of our fine-tuning, we use the Stable Audio model as the **baseline** and our fine-tuned music-specific model as the **experimental** model for comparison. Controlling for the same set of text prompts, we invoke both the baseline and fine-tuned models to generate music from a diverse selection of musical genres. In particular, we use the default parameters provided by Stable Audio Open 1.0: 20-second-long audio snippets, generated at a classifier-free guidance (CFG) scale parameter of 7 (controlling how much output generation follows the text prompt — higher values fit stronger to the prompt) [11].

Finally, we conduct evaluation of these outputs using a combination of both objective and subjective evaluation metrics. Objective metrics qualify the generated music based on values calculated from the audio data itself, while subjective metrics are determined by surveying listeners.

V. Results

We now discuss our results, first detailing the efficiency improvements acheived by using LoRA to fine-tune Stable Audio's diffusion model (as compared to fine-tuning without LoRA), then presenting the results of our conducted experimental setup. Finally, we will provide discussion and interpretation of these experimental results. Generated audio outputs are available in our repository.

A. LoRA Speedup

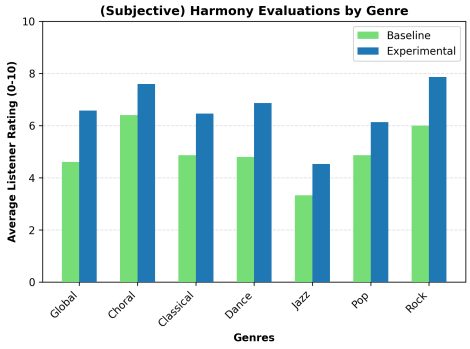
First, utilizing LoRA in our experimental model increased the efficiency of fine-tuning our audio diffusion by almost double. We tested this training speed-up by fine-tuning our baseline model without LoRA on the entire dataset for a few hours, for the sake of comparison with LoRA-aided fine-tuning. When using an A100 GPU and training on 5357 pieces of data, LoRA fine-tuning in the experimental model trained roughly 1.9375 epochs per hour, while the baseline model fine-tuned only 0.9 epochs per hour. Thus, by using LoRA, we fine-tuned the baseline model 115.278% faster than we would have performing a non-LoRA fine-tune.

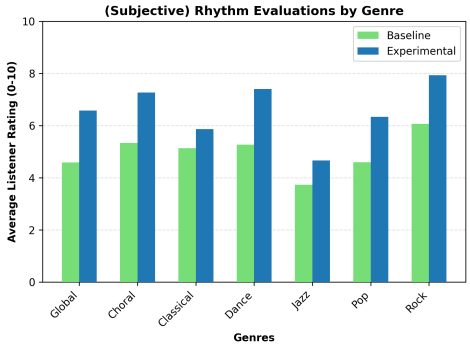
B. Music Generation

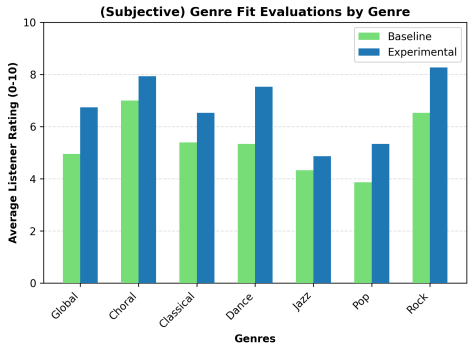
To evaluate our experimental model against the baseline, we generated music from a number of genres, namely: choral, classical, dance, jazz, pop, and rock. For each genre, we created prompts of increasing levels complexity, by increasing descriptiveness and word count, to build a diverse set of model outputs for each genre. These exact prompts can be found in our repository.

We control for this set of prompts in evaluating both the baseline and experimental models, ensuring all differences in model can be attributed to the models themselves (and stochasticity).

C. Subjective Evaluation:







We conduct subjective evaluation of the generated outputs through listener surveying. For each piece of music, listeners are asked to rate the following characteristics: **Harmony**—the quality of pitches, both in the melodies and chords of the music; **Rhythm**—the consistency and stability of beats within the music; and **Genre Fit**—The accuracy of the music to its categorized genre (ex. classical).

We conducted surveys on 5 listeners using all generated outputs across all genres and from both models. Averaging across genre and listener, we generated the following plots comparing the subjective evaluation of both models for each metric. Each plot additionally includes a global score averaged across all genres.

D. Objective Evaluation:

We additionally conduct objective evaluation on the generated outputs by computing a set of values for each piece of generated music directly from the audio files themselves. Specifically, we compute the following characteristics: **Harmonic Accuracy**—the percentage of pitches that belong to a standard scale, reporting for the best matching such scale; **Melodic Complexity**— the number of unique pitches in the music; and **Rhythmic Inconsistency**—the average degree to which pitches are off beat, expressed as a percentage of a single beat (quarter note).

We computed these values for all generated outputs, across all genres and for both models. Averaging across genre, we recorded these values in the following table. The table additionally includes a global score averaged across all genres:

TABLE I
OBJECTIVE METRICS

Objective Metrics	Experimental			Baseline		
	Harmonic Accuracy	Melodic Compl.	Rhythm Incons.	Harmonic Accuracy	Melodic Compl.	Rhythm Incons.
Global	91.4%	15.778	3.4%	91.2%	16.222	3.3%
Choral	81.3%	11	3.7%	93.3%	5.667	2.0%
Classical	87.7%	14.667	4.1%	85.3%	11.667	4.4%
Dance	98.3%	19	2.3%	88.0%	25.333	2.8%
Jazz	93.0%	18.667	3.6%	87.3%	20.333	4.3%
Pop	96.0%	15	3.4%	98.3%	16.667	3.0%
Rock	92.3%	16.333	3.0%	94.7%	17.667	3.2%

E. Interpretation and Discussion:

From our subjective survey results, we clearly observe that listeners show a preference towards the music produced by the experimental model across all metrics and genres. This is a clear indication that our fine-tuning results in a improvement in music quality to the human ear.

Interestingly, we note that both models demonstrate similar strengths and weaknesses when it comes to genre. Genres such as choral, dance, and rock score similarly high and genres such as jazz score similarly low. This is likely an indication that, while our fine-tuning leads to an improvement from the baseline model, both models maintain some degree of similarity. One reason for this could be that many of the baseline model weights are preserved when conducting LoRA fine-tuning, as only a small subset of weights are updated.

From our objective metrics, we note that both the baseline and experimental models demonstrate relatively high harmonic accuracy and low rhythmic inconsistency, indicating they are both produce outputs that are musically sound.

When approaching melodic complexity, we note that the baseline model demonstrates increased complexity for the genres dance, pop, and rock. When contextualized with our subjective results (as well our own intuition when listening), we note that this higher melodic complexity of the baseline in these genres appears to lead to more muddled and cacophonous music, while lower complexity in the experimental model appears to simpler, more digestible melody and harmony. This simplicity is intuitively important to the genres of dance, pop, and rock. We believe these results are an indication that our fine-tuning has captured this underlying concept.

For the genres of choral and classical music, the baseline model demonstrates noticeably low melodic complexity. Here, it appears the higher melodic complexity of our experimental model captures higher granularity in these genres.

Beyond these differences in melodic complexity, however, our objective evaluation results do not point towards the noticeable improvement in music quality reflected in our subjective evaluation. This is a reflection of a truth that we as musicians already believe - that music is more than just the right notes and rhythms - but rather their specific creative combination turned into art. In this sense, objective metrics alone appear to fall short in accurately evaluating the quality of music.

VI. CONCLUSION

This project successfully fine-tuned Stability AI's Stable Audio model for music generation using LoRA, achieving both high-quality audio outputs and computational efficiency. Our subjective evaluation highlights significant improvements in musical harmony, rhythm, and genre alignment, while objective metrics confirmed the model's ability to produce musically sound outputs. These findings underscore the potential of LoRA fine-tuning for domain-specific audio applications. Future work could explore more diverse datasets, extend model capabilities to longer compositions, and refine evaluation metrics to better capture the artistic nuances of music.

VII. Contribution Statement

Gabe's contributions: Model and Dataset Research; Data Refinement; Model Cloud Compute Research and Access; LoRA Fine-Tuning Implementation; Survey Interviews; and Results summarization and organization.

Jamin's contributions: Model and Dataset Research; Data Refinement; LoRA Fine-Tuning Implementation; Objective Evaluation; Subjective Evaluation; Survey Interviews; Writeup of Implementation and Results.

Liz's contributions: Model and Dataset Research; LoRA Fine-Tuning Training; Survey Interviews; Writeup of Abstract, Introduction, Related Work, Approach, Implementation, Conclusion, and aided in writing of Results.

VIII. Supplement

All of our code, data, and audio outputs can be found in this Google Drive. The "output audio" folder includes our LoRA fine-tuned experimental model outputs, as well as the non-fine-tuned Stable Audio model outputs on the same prompts.

References

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision." [Online]. Available: https://github.com/OpenAI/CLIP.

[2] B. Elizalde, S. Deshmukh, M. A. Ismail, and H. W. Microsoft, "Clap : Learning audio concepts from natural language supervision."

[3] Z. Evans, J. D. Parker, C. J. Carr, Z. Zukowski, J. Taylor, J. Pons, and S. Ai, "Stable audio open." [Online]. Available: https://github.com/Stability-AI/stable-audio-metrics

[4] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," 6 2021. [Online]. Available: http://arxiv.org/abs/2106.09685

[5] S. AI, "Stable audio v1.0," https://huggingface.co/stabilityai/stable-audio-open-1.0, 2024, accessed: 2024-12-11.

[6] O. Mogren, "C-rnn-gan: Continuous recurrent neural networks with gans for sequence generation," https://github.com/olofmogren/c-rnn-gan, 2018, accessed: 2024-12-11.

[7] Z. Xiong, W. Wang, J. Yu, Y. Lin, and Z. Wang, "A comprehensive survey for evaluation methodologies of ai-generated music," 8 2023. [Online]. Available: http://arxiv.org/abs/2308.13736

[8] Y.-C. Yeh, W.-Y. Hsiao, S. Fukayama, T. Kitahara, B. Genchel, H.-M. Liu, H.-W. Dong, Y. Chen, T. Leong, and Y.-H. Yang, "Automatic melody harmonization with triad chords: A comparative study," 1 2020. [Online]. Available: http://arxiv.org/abs/2001.02360

[9] A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi, M. Sharifi, N. Zeghidour, and C. Frank, "Musiclm: Generating music from text," 1 2023. [Online]. Available: http://arxiv.org/abs/2301.11325

[10] NeuralNotW0rk, "Loraw: Low-rank adaptation for audio waveforms," https://github.com/NeuralNotW0rk/LoRAW, 2024, accessed: 2024-12-11.

[11] D. Wang, "Stable diffusion guidance scale: Understanding the balance of creativity and fidelity," https://medium.com/@wangdk93/stablediffusion-guidance-scale-1822cfff7c9d, 2024, accessed: 2024-12-11.