MixingBuddy: A Multimodal LLM for Mix Critique and Advice

Pratham Vadhulas, Alexander Lerch

***Abstract*—this is the abstract. here is some more text. here is some more text.**

I. Introduction

Automatic Mixing, a key subfield of Music Informatics Research (MIR), aims to automate the complex and subjective task of music mixing. This area of study is pivotal to the modern music production and audio engineering market. To date, research in this field has made significant progress, largely by leveraging deep learning. Sophisticated models, such as U-Nets or generative frameworks, have been developed to make mixing systems accurate (predicting parameters that match professional mixes), controllable (allowing for high-level parameters to be set), and diverse (accomadating different genres and styles).

However, a critical limitation of these approaches is their "black box" nature. They can perform the mix, but they cannot explain their reasoning. The recent promise of large language models (LLMs) and multi-modal "agentic" systems introduces a new, necessary paradigm: explainability. We can now envision a tool that reasons about and discusses a mix.

This potential for co-creative, linguistic feedback brings us to our core research question: To what extent can an audio language model, when given a flawed mix, provide correct and useful advice?

II. Related Work

Early automatic mixing research centered on systems that captured expert knowledge through explicit mixing rules and heuristics. [1] developed an autonomous mixing system based on knowledge engineering principles. [2] employed probabilistic expert systems, a formal knowledge-based approach from early AI research, for automatic music production. Subsequent work explored machine learning techniques for instrument-specific effects, such as [3]'s approach to intelligent artificial reverberation application. [4] presented intelligent multitrack reverberation based on hinge-loss Markov random fields, demonstrating the application of statistical modeling tools to mixing tasks. [5] described a statistical approach to automated offline dynamic processing in the audio mastering process, further exemplifying the use of traditional machine learning methods. [6] introduced a framework that uses genetic optimization with timbral similarity measures. While these methods provided interpretable control and domain-specific optimization, they were limited in their ability to generalize across diverse musical styles and lacked the flexibility to adapt to unseen mixing scenarios.

The advent of deep learning brought significant advances in automatic mixing, with models capable of learning complex mappings from raw audio to mixing parameters. [7] introduced Wave-U-Net autoencoders for automatic mixing, demonstrating that end-to-end neural architectures could produce professional-quality mixes. [8] further advanced the field with a differentiable mixing console incorporating neural audio effects, enabling gradient-based optimization of mixing parameters. These deep learning approaches achieved notable success in terms of accuracy (matching professional mixes), controllability (allowing high-level parameter adjustment), and diversity (accommodating different genres and styles). However, a critical limitation of these systems is their "black box" nature: while they can perform the mix, they cannot explain their reasoning or provide linguistic feedback about mixing decisions.

Recognizing the need to bridge the semantic gap between audio processing and human understanding, researchers began exploring word-embedding approaches that link natural language descriptions to audio effect parameters. [9] demonstrated word embeddings for automatic equalization in audio mixing, while [10] developed Text2FX, which harnesses CLAP embeddings for text-guided audio effects. Early semantic mixing approaches [11] laid the groundwork for understanding how high-level, semantic knowledge could inform mixing decisions. These methods represented initial attempts to make mixing systems more interpretable and user-friendly by connecting linguistic descriptions to audio processing parameters.

Building on language-audio integration, recent work has explored prompt-driven interfaces that map natural language instructions directly to mixing tasks. [12] investigated whether large language models can predict audio effects parameters from natural language, while [13] developed SonicMaster, a controllable all-in-one music restoration and mastering system. [14] introduced MixAssist, an audio-language dataset for co-creative AI assistance in music mixing, demonstrating the potential for collaborative human-AI mixing workflows. These approaches represent an evolution toward more natural interaction paradigms, where users can express mixing intentions in natural language rather than manipulating low-level parameters.

Multimodal audio-language models have emerged as a powerful paradigm for combining audio understanding with language reasoning capabilities. One architectural approach, direct tokenization (also known as the unified approach), con-

Pratham Vadhulas is with the Georgia Institute of Technology, Atlanta, GA, USA (email: pvadhulas3@gatech.edu).

Alexander Lerch is with the Georgia Institute of Technology, Atlanta, GA, USA (email: alexander.lerch@gatech.edu).

verts raw audio into discrete tokens via audio codecs and extends the LLM vocabulary to include these audio tokens. This enables the language model to process audio and text within a unified framework. Key works in this direction include [15]'s AudioPaLM, [16]'s LauraGPT, and [17]'s SpeechGPT. The unified approach offers the advantage of treating audio and text as first-class citizens within the same model architecture, potentially enabling more seamless cross-modal reasoning.

An alternative architectural paradigm, the cascade (or feature extraction) approach, uses audio-specific encoders and decoders with the LLM serving as a central backbone. In this framework, audio is first encoded into feature representations that are then processed by the language model, which can generate text responses or guide audio generation. Examples include [18]'s M^2UGen and [19]'s Listen, Think, and Understand (LTU). The cascade approach allows for specialized audio processing while leveraging the reasoning capabilities of large language models, making it particularly relevant for tasks requiring both audio understanding and linguistic explanation, such as mix critique and advice.

Multimodal audio-language models show promise for explainable audio processing, but their application to mix critique and advice remains largely unexplored. Our work addresses this gap by leveraging multimodal audio-language models to provide linguistic feedback and reasoning about mixes. This positions our work as a step toward explainable, co-creative mixing systems that bridge the gap between automated processing and human understanding.

III. Methodology

To address this gap, we propose a multimodal audio-language model that takes a flawed mix as input and generates structured textual feedback identifying mixing flaws and suggesting corrective gain adjustments. Our approach focuses on relative gain relationships among multitrack stems, as gain balancing represents a fundamental and foundational aspect of mixing that directly addresses the core challenge of establishing proper balance between elements in a mix. This focus allows us to investigate whether the model can learn the relative nature of gain relationships where multiple valid solutions exist depending on the chosen anchor stem while providing a tractable starting point before extending to more complex mixing parameters such as equalization or dynamic processing. This methodology enables us to investigate our primary research question.

A. Dataset

For this work we will be using MUSDB18HQ as our ground truth dataset. We will be augmenting this dataset for our SFT training and DPO training. we know that the musdb18 dataset has 4 stems per track. these 4 stems are bass, drums, other and vocals. we can create flawed mixes by injecting errors into one of the stems and summing them. as a starting point we can choose one track and one type of mixing error that is gain. another thing we considered was that out of the 4 stems, other tends to be inconsistent, therefore we wont be using it as a target stem. Another thing to note, we know that gain balancing is a relative task, meaning there could be 2 solutions. for example, kick stem is set to be too loud. one solution is reducing kick. another solution is to increase all the other stems to match the kick. so theoretically, if we provide the model with an anchor stem, the model should be able to deduce which solution to suggest. again we didnt use other as an anchor in any of the sampels due to its inconsistency.

References

[1] E. Pérez-González and J. D. Reiss, "A knowledge-engineered autonomous mixing system," in *Audio Engineering Society Convention 135*, Oct. 2013.

[2] G. Bocko, M. F. Bocko, D. Headlam, J. Lundberg, and G. Ren, "Automatic music production system employing probabilistic expert systems," *journal of the audio engineering society*, 2010.

[3] E. Chourdakis and J. Reiss, "A machine-learning approach to application of intelligent artificial reverberation," *Journal of the Audio Engineering Society*, vol. 65, p. 56–65, Feb. 2017.

[4] A. L. Benito and J. D. Reiss, "Intelligent Multitrack Reverberation Based on Hinge-Loss Markov Random Fields," in *Proceedings of the 2017 AES International Conference on Semantic Audio*, June 2017.

[5] M. Hilsamer and S. Herzog, "A statistical approach to automated offline dynamic processing in the audio mastering process," in *Proc. of the 17th Int. Conference on Digital Audio Effects (DAFx-14)*, September 2014.

[6] B. Kolasinski, "A framework for automatic mixing using timbral similarity measures and genetic optimization," vol. 3, 05 2008.

[7] E. Chourdakis and J. D. Reiss, "Automatic music signal mixing system based on one-dimensional wave-u-net autoencoders," *EURASIP Journal on Audio, Speech, and Music Processing*, 2022.

[8] C. J. Steinmetz, J. Pons, S. Pascual, and J. Serrà, "Automatic multitrack mixing with a differentiable mixing console of neural audio effects," Oct. 2020.

[9] S. Venkatesh, D. Moffat, and E. R. Miranda, "Word embeddings for automatic equalization in audio mixing," *Journal of the Audio Engineering Society*, vol. 70, p. 753–763, Nov. 2022.

[10] A. Chu, P. O'Reilly, J. Barnett, and B. Pardo, "Text2fx: Harnessing clap embeddings for text-guided audio effects," Feb. 2025.

[11] E. Chourdakis and J. Reiss, "A semantic approach to autonomous mixing," 2016.

[12] S. Doh, J. Koo, M. A. Martínez-Ramírez, W.-H. Liao, J. Nam, and Y. Mitsufuji, "Can large language models predict audio effects parameters from natural language?" Jul. 2025.

[13] J. Melechovsky, A. Mehrish, and D. Herremans, "Sonicmaster: Towards controllable all-in-one music restoration and mastering," Aug. 2025.

[14] M. P. Clemens and A. Marasovic, "Mixassist: An audio-language dataset for co-creative AI assistance in music mixing," 2025.

[15] P. K. Rubenstein, C. Asawaroengchai, D. D. Nguyen, and et al., "Audiopalm: A large language model that can speak and listen," Jun. 2023.

[16] Z. Du, J. Wang, Q. Chen, Y. Chu, Z. Gao, and et al., "Lauragpt: Listen, attend, understand, and regenerate audio with gpt," Jul. 2024.

[17] D. Zhang, S. Li, X. Zhang, J. Zhan, P. Wang, Y. Zhou, and X. Qiu, "Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities," May 2023.

[18] S. Liu, A. S. Hussain, Q. Wu, C. Sun, and Y. Shan, "M^2ugen: Multi-modal music understanding and generation with the power of large language models," Dec. 2024.

[19] Y. Gong, H. Luo, A. H. Liu, L. Karlinsky, and J. Glass, "Listen, think, and understand," Feb. 2024.