Differential Multi-Track Gain Analysis for AI Mixing Assistance

Pratham Vadhulas

September 16, 2025

1 Research Statement

Current Audio-Language Models (ALMs) excel at descriptive audio tasks, but tasks like music mixing require comparative analysis and relational understanding that go beyond simple description. This research proposes a novel differential analysis framework that conditions ALMs on multi-track audio to provide more specific and actionable gain-balancing guidance. Instead of analyzing single mixes, our approach trains models to compare unbalanced and balanced multi-track sets, learning the causal relationships between specific gain adjustments and perceived mix improvements. To achieve this, the study will leverage multiple datasets: an augmented version of MUSDB18 with human consensus gain values, "The Mix Evaluation Dataset" which contains mixing decisions from human engineers, and its corresponding "MixParams" metadata dataset available on Hugging Face. Performance will be evaluated with a comprehensive methodology, including automated LLM-as-a-Judge assessments and human preference studies. The primary research question investigates whether this differential approach improves the technical specificity and user-perceived helpfulness of AI-generated gain-balancing advice, providing a foundation for more sophisticated AI mixing assistants that can reason about the relational properties of multi-track audio.

1.1 Research Questions

1.1.1 Primary Research Question

To what extent does a differential analysis framework, conditioned on multi-track audio, improve the technical specificity and user-perceived helpfulness of AI-generated gain-balancing advice compared to a traditional single-mix advisory model?

1.1.2 Secondary Research Questions

1. How effectively can an Audio-Language Model be fine-tuned on a synthetic dataset of 'problem' and 'solution' multi-track sets to learn the causal relationship between specific gain adjustments and perceived improvements in mix balance?

2. What is an effective architectural approach for representing and comparing two parallel sets of multi-track stems to enable an LLM to reason about their relative gain differences?

3. For the specific task of evaluating gain-balancing advice, what is the correlation between automated evaluation (i.e., LLM-as-a-Judge rankings) and subjective human preference judgments?

1.2 Scope and Limitations

This research is specifically focused on the task of generating textual advice for gain parameter adjustments in multi-track audio. The core of the study is the development of a novel differential analysis framework trained on the MUSDB18, "The Mix Evaluation Dataset," and "MixParams" datasets. The model's output will be advisory text, not direct, continuous parameter predictions (e.g., -2.5dB). The investigation deliberately excludes other mixing parameters such as equalization (EQ), dynamic range compression, and spatial effects to maintain a focused scope. Furthermore, the proposed system is designed for offline analysis and is not intended for real-time, interactive applications. The validity of this work relies on the assumption that the human-derived values in the chosen datasets represent a perceptually valid ground truth for a well-balanced mix and that the datasets are of sufficient quality for the task.

2 Motivation

2.1 Research Gap

A significant gap exists in the current automatic mixing research landscape. The field is largely dominated by black-box, deep learning models that, while achieving impressive results, offer little to no explainability into their decision-making process. This contrasts with other creative domains like 3D modeling and image editing, which have seen the rise of end-to-end agentic systems that facilitate user interaction and provide transparency.

Furthermore, while generative audio models have gained popularity, their output is often of limited value to professional mixing and mastering engineers, who require precise control over mixing parameters rather than a final, uneditable audio file.

To bridge this gap, future research must move towards models with better audio and contextual understanding. Exploring techniques like audio-conditioned chain-of-thought reasoning, combined with state-of-the-art automatic mixing models, will be crucial. This direction will not only improve performance but also foster trust and adoption by professionals by providing explainable and controllable systems.

2.2 Research Impact

The development of explainable and interactive automatic mixing systems could have a profound impact on the music production industry. Digital Audio Workstations (DAWs) could integrate features where users can converse with an intelligent mixing agent to receive advice, ask specific questions, and collaboratively edit mixes. This paradigm shifts the focus from replacing engineers with end-to-end generative systems to empowering musicians and producers with powerful, intuitive tools. By fostering a more interactive and transparent mixing process, this research can enhance creativity and streamline workflows for audio professionals and hobbyists alike.

3 Related Work

Addressing the gaps in current automatic mixing systems namely the lack of explainability, interactivity, and user control requires a conceptual shift away from fully autonomous "black box" solutions. The emerging field of "Co-Creative AI" provides a compelling framework for this shift, envisioning AI not as a replacement for human creativity, but as an intelligent partner that augments the artistic workflow [1, 2]. This collaborative paradigm is particularly resonant in a field as nuanced and subjective as audio mixing, where artistic intent is paramount. Recent work has demonstrated the potential of co-creative systems in music production, showing positive user adoption and acceptance when AI tools provide appropriate levels of control and collaboration [3, 4]. This section reviews the evolution of co-creative systems and the underlying technologies that make a language-driven, interactive mixing agent feasible.

3.1 Paradigms of Co-Creative AI in Audio

Within the landscape of co-creative systems for audio engineering, early paradigms focused on automating the mixing process. These automated mixers often functioned as **black box systems**, which take raw multitrack audio as input to produce a fully mixed output with minimal user intervention [5, 6]. While some of these systems provide a degree of user control through high-level parameters, they fundamentally abstract the underlying process, which limits the fine-grained control required in professional workflows. In contrast, a more recent and promising paradigm is the **Language Bridge**, where natural language serves as the primary interface for interaction. This approach allows users to articulate creative intent in descriptive terms, and our research is situated within this paradigm. Recent work has begun exploring this direction across multiple dimensions: systems like MixAssist demonstrate the potential for audio-language models to provide contextual mixing advice through natural language dialogue [7]; Text2FX leverages CLAP embeddings to control audio effects through open-vocabulary natural language prompts [8]; and LLM2Fx shows that Large Language Models can predict audio effect parameters directly from textual descriptions in a zero-shot manner [9]. Additionally, research has explored speech recognition for mixer control [10], natural language interfaces for audio production tools [11], and word embeddings for automatic equalization [12].

3.2 The Evolution of Language-Driven Models

The journey towards sophisticated language-driven models has been incremental, progressing through several key phases that have transformed how we interact with and control various modalities through natural language. It began with **task-specific supervised models** for applications like audio tagging, which classified sounds into predefined categories [13]. A subsequent shift towards **representation learning** aimed to create more general-purpose audio embeddings to capture richer semantic information [14]. The emergence of **contrastive dual-encoder models** like CLAP marked a significant advancement, learning audio concepts from natural language supervision and enabling zero-shot audio classification and retrieval [15, 16]. The rise of deep **generative models** like GANs and VAEs enabled audio synthesis and transformation, though precise control remained a challenge [17, 18]. More recently, **generative and**

diffusion-based multimodal models have emerged, leveraging the power of diffusion models and large language models for text-to-audio generation [19, 20]. Most significantly, the development of **Large Audio-Language Models (LALMs)** and instruction-following systems has been revolutionary, with models like AudioLM demonstrating language modeling approaches to audio generation [21]. The profound language understanding of these models, when combined with the ability to process audio, forms the technological cornerstone of the Language Bridge paradigm, making it possible to connect linguistic intent directly to audio manipulation [22].

3.3 Multimodal Large Language Models

The development of Multimodal Large Language Models (MM-LLMs) represents a significant advancement in AI capabilities, enabling models to process and understand multiple modalities simultaneously. Several architectural paradigms have emerged for creating these powerful systems:

Unified Multimodal Models are trained from the ground up on vast datasets spanning multiple modalities, allowing for deep integration of information across different data types [23]. These models, such as Unified-IO, demonstrate the potential for truly unified understanding across vision, language, and audio domains.

Modality Interface Architectures involve augmenting pre-trained, text-only LLMs with specialized encoders for other modalities. This approach allows LLMs to perceive and process new types of information without altering their core architecture. Notable examples include Qwen-Audio, which provides universal audio understanding capabilities across over 30 tasks [24], and SALMONN, which integrates speech and audio encoders with LLMs to achieve generic hearing abilities [25].

Any-to-Any Multimodal Systems represent the cutting edge, enabling models to both perceive and generate content across multiple modalities. NExT-GPT demonstrates this capability, connecting LLMs with multimodal adaptors and diffusion decoders to handle arbitrary combinations of text, image, video, and audio [26]. Similarly, Audiobox provides unified audio generation capabilities across speech, sound, and music through natural language prompts [27].

Instruction-Following Multimodal Models focus on following complex instructions across modalities. Macaw-LLM seamlessly integrates visual, audio, and textual information for multi-turn dialogue scenarios [28], while PandaGPT demonstrates emergent cross-modal behaviors through instruction-following capabilities [29].

3.4 Fine-Tuning Methods for Audio Tasks

The application of fine-tuning to MM-LLMs has unlocked a wide range of capabilities across various audio domains. This section provides an overview of the state-of-the-art methods for general audio sounds, music, and speech.

3.4.1 General Audio Sounds

Audio Understanding Models such as LTU [30], SALMONN [25], Qwen-Audio [24], and UNIFIED-IO 2 [31] leverage LLMs as their backbone for analyzing and interpreting diverse

environmental sounds. Additionally, AudioGPT [32] and HuggingGPT [**?**] function as intelligent interfaces that coordinate various tools for audio understanding tasks. Furthermore, recent work has enhanced automated audio captioning by integrating pretrained models with LLMs [33].

Audio Generation Notable models in audio generation include TANGO [34], Make-an-Audio 2 [35], WavJourney [36], AudioLM [21], Audiobox [27], and UniAudio [37]. These approaches utilize a variety of techniques, such as text embedders (e.g., FLAN-T5 in TANGO), latent diffusion models, LLM agents for integrating audio models (WavJourney), discrete audio tokenization (AudioLM), LLMs for data construction and flow-matching (Audiobox), and unified sequence tokenization for various audio types (UniAudio).

3.4.2 Music

Music Understanding In the music domain, models like Music Understanding LLaMA (MU-LLaMA) [38], LLARK [39], MusicAgent [40], LyricWhiz [41], and ChatMusician [42] are employed to analyze detailed music features, leverage refined annotations, automate tasks, and improve lyric transcription.

Music Generation Music generation methods include MusicLM [43], Jukebox [44], MusicGen [45], Music ControlNet [33], M2UGen [46], ChatMusician [42], and SongComposer [47]. These often use Transformer architectures for conditional music generation, compress raw audio into discrete codes (Jukebox), incorporate LLMs as text embedders (MusicGen, Music ControlNet), combine LLMs with other pretrained models (M2UGen), or intrinsically generate symbolic music (ChatMusician, SongComposer).

Music Editing Loop Copilot [48] combines LLMs with specialized AI music models to facilitate conversational, collaborative music loop creation and editing.

3.4.3 Speech

Speech Understanding Key contributions in speech understanding come from SpeechGPT [49], AudioPaLM [50], Speech-LLaMA [51], and recent works that utilize LLMs as structural backbones to process spoken language, support multimodal content, transfer inter-modal knowledge, and improve Automatic Speech Recognition (ASR) accuracy through in-context learning or specialized connector structures [52, 53].

Speech Generation VALL-E [54] uses a neural codec language model to reframe text-to-speech (TTS) as a conditional language modeling task. Other approaches integrate LLaMA/OPT with VALL-E [55], and LauraGPT [56] is a unified GPT model for speech recognition, translation, and TTS. Additionally, some research investigates word surprisal to improve speech synthesis prosody [57].

3.5 How LLMs are Utilized in Audio Tasks

The integration of LLMs in audio tasks can be categorized into several key approaches:

- **LLMs as Backbone:** Pre-trained LLMs (e.g., LLaMA) are used as the central architecture, either with modality-specific encoders/decoders (cascade approach) or by tokenizing raw audio into discrete tokens for direct LLM input (unified approach).
- **LLMs as Conditioner:** LLMs encode text prompts into embeddings that condition the audio generation process.
- **LLMs as Labeller:** LLMs are employed to convert class labels from large audio datasets into full-sentence audio descriptions or captions, often utilizing self-instruction techniques.
- **LLMs as Agent:** LLMs act as controllers, interfacing with and orchestrating various external tools to accomplish diverse audio tasks.
- **LLMs Inspired Backbone:** This approach discretizes audio into tokens for next-token prediction, aiming for LLM-like emergent capabilities in audio.

Furthermore, tool-augmented multimodal agents like ControlLLM [58], ModelScope-Agent [59], and HuggingGPT [**?**] can generate speech and music by invoking specialized audio tools. NExT-GPT also provides a framework that supports mixed inputs and outputs including audio, with diffusion models attached to the LLM [26].

Our research falls within the domain of fine-tuning Multimodal LLMs, specifically investigating how these models can be adapted to serve as intuitive and effective co-creative partners in the complex task of audio mixing. We aim to leverage a modality interface architecture, enabling a powerful LLM to understand and act upon instructions related to multitrack audio.

4 Proposed Method

4.1 Overall Architecture

The proposed system will employ a multi-stage approach combining:

- **Musical Analysis**: Understanding of song structure, genre, and musical relationships
- **Context-Aware Processing**: Track-level analysis considering the full mix context
- **Adaptive Parameter Prediction**: Machine learning models for mixing parameter estimation
- **Iterative Refinement**: Feedback mechanisms for continuous improvement

4.2 Key Components

4.2.1 Musical Context Analysis

- Genre classification and style analysis
- Song structure detection (verse, chorus, bridge)
- Instrument role identification and importance ranking
- Harmonic and rhythmic analysis

4.2.2 Multi-Track Feature Extraction

- Spectral features (MFCCs, spectral centroid, rolloff)
- Temporal features (RMS, peak levels, attack/decay)
- Perceptual features (loudness, brightness, roughness)
- Cross-track correlation and masking analysis

4.2.3 Neural Network Architecture

The proposed model will use:

- **Encoder-Decoder Structure**: For understanding track relationships
- **Attention Mechanisms**: To focus on relevant track interactions
- **Multi-Task Learning**: Simultaneous prediction of multiple mixing parameters
- **Transfer Learning**: Leveraging pre-trained models for musical understanding

4.3 Training Strategy

- **Dataset**: Large-scale collection of professionally mixed tracks with parameter annotations
- **Data Augmentation**: Synthetic variations of existing mixes
- **Curriculum Learning**: Progressive training from simple to complex mixing scenarios
- **Multi-Objective Optimization**: Balancing technical quality with musical appropriateness

4.4 Feasibility Considerations

The proposed approach is feasible because:

- Existing datasets of mixed tracks are available for training
- Modern deep learning frameworks can handle the computational requirements
- The modular design allows for incremental development and testing
- Industry partnerships can provide access to professional mixing data

5 Proposed Evaluation

5.1 Evaluation Framework

A comprehensive evaluation strategy will be developed to assess both technical performance and perceptual quality of the automatic mixing system.

5.2 Objective Metrics

- **Technical Accuracy**: Comparison of predicted parameters with ground truth values
- **Spectral Analysis**: Frequency response, dynamic range, and harmonic content
- **Level Balancing**: RMS and peak level distributions across tracks
- **Spatial Characteristics**: Stereo width, panning accuracy, and spatial coherence

5.3 Subjective Evaluation

- **Listening Tests**: A/B comparisons with professional mixes
- **Expert Evaluation**: Assessment by professional audio engineers
- **Genre-Specific Tests**: Evaluation across different musical styles
- **Long-term Listening**: Assessment of mix fatigue and musicality

5.4 Comparative Analysis

The system will be compared against:

- Current state-of-the-art automatic mixing systems
- Rule-based mixing approaches
- Human-engineered mixes (both amateur and professional)
- Baseline systems (e.g., simple level balancing)

5.5 Evaluation Datasets

- **Professional Mixes**: High-quality reference mixes from various genres
- **Amateur Productions**: User-generated content for generalization testing
- **Synthetic Data**: Generated mixes for controlled evaluation scenarios
- **Cross-Genre Data**: Diverse musical styles for robustness testing

5.6 Statistical Analysis

- Significance testing for subjective evaluations
- Correlation analysis between objective and subjective metrics
- Error analysis and failure case identification
- Performance analysis across different musical contexts

6 Novelty of Proposed Work

6.1 Advancements in Problem Formulation

- **Musical Context Integration**: First systematic approach to incorporating musical understanding into automatic mixing
- **Multi-Scale Analysis**: Novel framework for analyzing both individual tracks and full mix context simultaneously
- **Genre-Adaptive Processing**: Adaptive system that adjusts mixing strategies based on musical genre and style

6.2 Methodological Innovations

- **Attention-Based Architecture**: Novel use of attention mechanisms for track relationship modeling
- **Multi-Task Learning Framework**: Simultaneous optimization of multiple mixing parameters with shared representations
- **Iterative Refinement**: Feedback mechanisms that allow the system to improve its output over multiple iterations
- **Interpretable AI**: Development of explainable mixing decisions for human understanding and control

6.3 System Architecture Contributions

- **Modular Design**: Flexible architecture that allows for component-wise evaluation and improvement
- **Real-Time Capability**: Efficient processing pipeline suitable for interactive applications
- **Scalability**: System design that can handle varying numbers of tracks and complexity levels
- **Integration Framework**: Seamless integration with existing digital audio workstations (DAWs)

6.4 Evaluation Methodology Advances

- **Comprehensive Evaluation Framework**: Novel combination of objective and subjective evaluation methods
- **Genre-Specific Metrics**: Development of evaluation criteria tailored to different musical styles
- **Longitudinal Studies**: Long-term evaluation of mixing quality and user satisfaction
- **Cross-Cultural Validation**: Evaluation across different musical traditions and cultural contexts

6.5 Impact on State-of-the-Art

This work advances the field by:

- Moving beyond simple parameter prediction to musical understanding
- Establishing new benchmarks for automatic mixing evaluation
- Creating reusable components for future research in computational audio
- Bridging the gap between technical audio processing and musical artistry

7 Required Resources

7.1 Computational Resources

- **High-Performance Computing**: Access to GPU clusters for training large neural networks
- **Storage**: Large-scale storage for audio datasets (estimated 10+ TB)
- **Processing Power**: Multi-core workstations for real-time processing and development
- **Cloud Computing**: Access to cloud platforms for scalable training and evaluation

7.2 Software and Tools

- **Deep Learning Frameworks**: PyTorch, TensorFlow, or similar for model development
- **Audio Processing Libraries**: LibROSA, Essentia, or similar for audio analysis
- **Digital Audio Workstations**: Professional DAWs for reference and testing
- **Evaluation Tools**: Custom software for objective and subjective evaluation

7.3 Datasets and Data

- **Professional Mixes**: Access to high-quality mixed tracks with parameter annotations
- **Multi-Track Recordings**: Raw tracks for training and testing
- **Metadata**: Genre labels, mixing notes, and production information
- **Reference Standards**: Industry-standard reference tracks for evaluation

7.4 Human Resources

- **Audio Engineers**: Professional mixing engineers for consultation and evaluation
- **Musicians**: Artists for providing diverse musical content
- **Research Collaborators**: Experts in machine learning, audio processing, and music cognition
- **User Testers**: Both professional and amateur users for system evaluation

7.5 Equipment and Facilities

- **Recording Studio**: Access to professional recording facilities
- **Monitoring Systems**: High-quality audio monitoring for evaluation
- **Acoustic Treatment**: Proper acoustic environment for listening tests
- **Research Lab**: Dedicated space for development and testing

7.6 Budget Considerations

- **Computing Costs**: GPU rental and cloud computing expenses
- **Data Licensing**: Costs for accessing commercial music datasets
- **Equipment**: Audio equipment and software licenses
- **Personnel**: Compensation for expert consultants and testers

8 Deliverables

8.1 Research Publications

- **Conference Papers**: 2-3 papers at top-tier conferences (ISMIR, ICASSP, AES)
- **Journal Articles**: 1-2 papers in high-impact journals (JASA, IEEE TASLP)
- **Workshop Presentations**: Presentations at relevant workshops and symposiums
- **Technical Reports**: Detailed technical documentation of methods and results

8.2 Software and Code

- **Open-Source Implementation**: Complete source code of the automatic mixing system
- **API and SDK**: Software development kit for integration with existing tools
- **Plug-in Development**: VST/AU plugins for popular digital audio workstations
- **Web Application**: Browser-based interface for automatic mixing

8.3 Datasets and Resources

- **Training Datasets**: Curated datasets of mixed tracks with annotations
- **Evaluation Benchmarks**: Standardized test sets for system comparison
- **Pre-trained Models**: Trained models ready for use and further development
- **Documentation**: Comprehensive documentation for datasets and models

8.4 Evaluation Tools

- **Evaluation Framework**: Software tools for objective and subjective evaluation
- **Benchmarking Suite**: Automated testing and comparison tools
- **Visualization Tools**: Software for analyzing and visualizing mixing decisions
- **User Study Materials**: Protocols and materials for human evaluation studies

8.5 Documentation and Tutorials

- **User Manuals**: Comprehensive guides for end users
- **Developer Documentation**: Technical documentation for researchers and developers
- **Tutorial Videos**: Educational content demonstrating system capabilities
- **Best Practices Guide**: Recommendations for optimal system usage

8.6 Intellectual Property

- **Patents**: Novel algorithms and methods (if applicable)
- **Open Source Licenses**: Appropriate licensing for public release
- **Commercial Licensing**: Options for commercial use and integration
- **Research Agreements**: Collaboration agreements with industry partners

8.7 Dissemination

- **Conference Presentations**: Oral and poster presentations
- **Demonstrations**: Live demonstrations at conferences and workshops
- **Media Coverage**: Press releases and media outreach
- **Community Engagement**: Participation in relevant online communities and forums

9 Timeline

9.1 Phase 1: Foundation and Data Collection (Weeks 1-8)

- **Weeks 1-2**: Literature review and system architecture design
- **Weeks 3-4**: Dataset collection and preprocessing pipeline development
- **Weeks 5-6**: Feature extraction and musical analysis framework
- **Weeks 7-8**: Baseline system implementation and initial testing

9.2 Phase 2: Core System Development (Weeks 9-20)

- **Weeks 9-12**: Neural network architecture design and implementation
- **Weeks 13-16**: Training pipeline development and initial model training
- **Weeks 17-20**: System integration and real-time processing optimization

9.3 Phase 3: Evaluation and Refinement (Weeks 21-32)

- **Weeks 21-24**: Comprehensive evaluation framework development
- **Weeks 25-28**: Objective and subjective evaluation studies
- **Weeks 29-32**: System refinement based on evaluation results

9.4 Phase 4: Advanced Features and Optimization (Weeks 33-44)

- **Weeks 33-36**: Genre-adaptive processing and advanced features
- **Weeks 37-40**: User interface development and usability testing
- **Weeks 41-44**: Performance optimization and scalability improvements

9.5 Phase 5: Validation and Dissemination (Weeks 45-52)

- **Weeks 45-48**: Large-scale validation studies and user testing
- **Weeks 49-52**: Paper writing, software release, and dissemination

9.6 Milestones and Deliverables

- **Week 8**: Baseline system and initial results
- **Week 16**: First working prototype
- **Week 24**: Evaluation framework and initial evaluation results
- **Week 32**: Refined system with improved performance
- **Week 40**: Complete system with user interface
- **Week 48**: Final validation results
- **Week 52**: Final deliverables and publications

9.7 Risk Mitigation

- **Data Availability**: Alternative datasets and synthetic data generation
- **Technical Challenges**: Incremental development with fallback options
- **Evaluation Difficulties**: Multiple evaluation approaches and expert consultation
- **Timeline Delays**: Buffer time built into each phase

References

[1] R. Louie, A. Coenen, C. Z. Huang, M. Terry, and C. J. Cai, "Novice-ai music co-creation via ai-steering tools for deep generative models," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, (Honolulu HI USA), p. 1–13, ACM, Apr. 2020.

[2] A. Tsiros and A. Palladini, "Towards a human-centric design framework for ai assisted music production," June 2020.

[3] R. Bougueng Tchemeube, J. Ens, C. Plut, P. Pasquier, M. Safi, Y. Grabit, and J.-B. Rolland, "Evaluating human-ai interaction via usability, user experience and acceptance measures for mmm-c: A creative ai system for music composition," in *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, (Macau, SAR China), p. 5769–5778, International Joint Conferences on Artificial Intelligence Organization, Aug. 2023.

[4] S. S. Vanka, M. Safi, J.-B. Rolland, and G. Fazekas, "Adoption of ai technology in the music mixing workflow: An investigation," Sept. 2023. arXiv:2304.03407 [cs].

[5] C. J. Steinmetz, J. Pons, S. Pascual, and J. Serrà, "Automatic multitrack mixing with a differentiable mixing console of neural audio effects," Oct. 2020. arXiv:2010.10291 [eess].

[6] M. A. Martínez-Ramírez, W.-H. Liao, G. Fabbro, S. Uhlich, C. Nagashima, and Y. Mitsufuji, "Automatic music mixing with deep learning and out-of-domain data," Aug. 2022. arXiv:2208.11428 [eess].

[7] M. Clemens and A. Marasović, "Mixassist: An audio-language dataset for co-creative ai assistance in music mixing," July 2025. arXiv:2507.06329 [cs].

[8] A. Chu, P. O'Reilly, J. Barnett, and B. Pardo, "Text2fx: Harnessing clap embeddings for text-guided audio effects," Feb. 2025. arXiv:2409.18847 [eess].

[9] 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?," July 2025. arXiv:2505.20770 [cs].

[10] S.-C. Lai, Y.-H. Hung, Y.-C. Zhu, S.-T. Wang, M.-H. Sheu, and W.-H. Juang, "A low-cost smart digital mixer system based on speech recognition," *Electronics*, vol. 11, p. 604, Feb. 2022.

[11] B. Pardo, M. Cartwright, P. Seetharaman, and B. Kim, "Learning to build natural audio production interfaces," *Arts*, vol. 8, p. 110, Aug. 2019.

[12] 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.

[13] Q. Kong, Y. Xu, and M. D. Plumbley, "Attention-based deep multiple instance learning for weakly supervised audio tagging," in *2018 26th European Signal Processing Conference (EUSIPCO)*, pp. 111–115, IEEE, 2018.

[14] K. Choi, J. Lee, and J. Nam, "Content-based music similarity with deep representation learning," in *Proceedings of the 18th International Society for Music Information Retrieval Conference (ISMIR 2017)*, 2017.

[15] B. Elizalde, S. Deshmukh, M. A. Ismail, and H. Wang, "Clap learning audio concepts from natural language supervision," in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 1–5, June 2023.

[16] A. S. Koepke, A.-M. Oncescu, J. F. Henriques, Z. Akata, and S. Albanie, "Audio retrieval with natural language queries: A benchmark study," *IEEE Transactions on Multimedia*, vol. 25, p. 2675–2685, 2023.

[17] C. Donahue, J. McAuley, and M. Puckette, "Wavegan: A gan for raw audio synthesis," in *International Conference on Learning Representations*, 2018.

[18] J. Engel, A. Roberts, S. Dieleman, D. Askew, S. Oore, and D. Eck, "Disentangled representations of musical timbre with gans and vaes," *arXiv preprint arXiv:1911.08323*, 2019.

[19] H. Liu, Y. Yuan, X. Liu, X. Mei, Q. Kong, Q. Tian, Y. Wang, W. Wang, Y. Wang, and M. D. Plumbley, "Audioldm 2: Learning holistic audio generation with self-supervised pretraining," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, p. 2871–2883, 2024.

[20] J. Xue, Y. Deng, Y. Gao, and Y. Li, "Auffusion: Leveraging the power of diffusion and large language models for text-to-audio generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, p. 4700–4712, 2024.

[21] Z. Borsos, R. Marinier, D. Vincent, E. Kharitonov, O. Pietquin, M. Sharifi, D. Roblek, O. Teboul, D. Grangier, M. Tagliasacchi, and N. Zeghidour, "Audiolm: A language modeling approach to audio generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, p. 2523–2533, 2023.

[22] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.

[23] J. Lu, C. Clark, R. Zellers, R. Mottaghi, and A. Kembhavi, "Unified-io: A unified model for vision, language, and audio," in *European Conference on Computer Vision*, pp. 525–542, Springer, 2022.

[24] Y. Chu, J. Xu, X. Zhou, Q. Yang, S. Zhang, Z. Yan, C. Zhou, and J. Zhou, "Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models," Dec. 2023. arXiv:2311.07919 [eess].

[25] C. Tang, W. Yu, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, Z. Ma, and C. Zhang, "Salmonn: Towards generic hearing abilities for large language models," Apr. 2024. arXiv:2310.13289 [cs].

[26] S. Wu, H. Fei, L. Qu, W. Ji, and T.-S. Chua, "Next-gpt: Any-to-any multimodal llm," June 2024.

[27] A. Vyas, B. Shi, M. Le, A. Tjandra, Y.-C. Wu, B. Guo, J. Zhang, X. Zhang, R. Adkins, W. Ngan, J. Wang, I. Cruz, B. Akula, A. Akinyemi, B. Ellis, R. Moritz, Y. Yungster, A. Rakotoarison, L. Tan, C. Summers, C. Wood, J. Lane, M. Williamson, and W.-N. Hsu, "Audiobox: Unified audio generation with natural language prompts," Dec. 2023. arXiv:2312.15821 [cs].

[28] C. Lyu, M. Wu, L. Wang, X. Huang, B. Liu, Z. Du, S. Shi, and Z. Tu, "Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration," June 2023. arXiv:2306.09093 [cs].

[29] Y. Su, T. Lan, H. Li, J. Xu, Y. Wang, and D. Cai, "Pandagpt: One model to instruction-follow them all," May 2023. arXiv:2305.16355 [cs].

[30] Y. Gong, H. Luo, A. H. Liu, L. Karlinsky, and J. Glass, "Listen, think, and understand," Feb. 2024. arXiv:2305.10790 [eess].

[31] J. Lu, C. Clark, S. Lee, Z. Zhang, S. Khosla, R. Marten, D. Hoiem, and A. Kembhavi, "Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action," p. 26439–26455, 2024.

[32] R. Huang, M. Li, D. Yang, J. Shi, X. Chang, Z. Ye, Y. Wu, Z. Hong, J. Huang, J. Liu, Y. Ren, Z. Zhao, and S. Watanabe, "Audiogpt: Understanding and generating speech, music, sound, and talking head," Apr. 2023. arXiv:2304.12995 [cs].

[33] S.-L. Wu, C. Donahue, S. Watanabe, and N. J. Bryan, "Music controlnet: Multiple time-varying controls for music generation," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, p. 2692–2703, 2024.

[34] D. Ghosal, N. Majumder, A. Mehrish, and S. Poria, "Text-to-audio generation using instruction-tuned llm and latent diffusion model," May 2023. arXiv:2304.13731 [eess].

[35] J. Huang, Y. Ren, R. Huang, D. Yang, Z. Ye, C. Zhang, J. Liu, X. Yin, Z. Ma, and Z. Zhao, "Make-an-audio 2: Temporal-enhanced text-to-audio generation," May 2023. arXiv:2305.18474 [cs].

[36] X. Liu, Z. Zhu, H. Liu, Y. Yuan, Q. Huang, M. Cui, J. Liang, Y. Cao, Q. Kong, M. D. Plumbley, and W. Wang, "Wavjourney: Compositional audio creation with large language models," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 33, p. 2830–2844, 2025.

[37] D. Yang, J. Tian, X. Tan, R. Huang, S. Liu, X. Chang, J. Shi, S. Zhao, J. Bian, Z. Zhao, X. Wu, and H. Meng, "Uniaudio: An audio foundation model toward universal audio generation," Dec. 2024. arXiv:2310.00704 [cs].

[38] S. Liu, A. S. Hussain, C. Sun, and Y. Shan, "Music understanding llama: Advancing text-to-music generation with question answering and captioning," in *ICASSP 2024 - 2024 IEEE*

International Conference on Acoustics, Speech and Signal Processing (ICASSP), p. 286–290, Apr. 2024.

[39] J. Gardner, S. Durand, D. Stoller, and R. M. Bittner, "Llark: A multimodal instruction-following language model for music," June 2024. arXiv:2310.07160 [cs].

[40] D. Yu, K. Song, P. Lu, T. He, X. Tan, W. Ye, S. Zhang, and J. Bian, "Musicagent: An ai agent for music understanding and generation with large language models," Oct. 2023. arXiv:2310.11954 [cs].

[41] L. Zhuo, R. Yuan, J. Pan, Y. Ma, Y. LI, G. Zhang, S. Liu, R. Dannenberg, J. Fu, C. Lin, E. Benetos, W. Xue, and Y. Guo, "Lyricwhiz: Robust multilingual zero-shot lyrics transcription by whispering to chatgpt," July 2024. arXiv:2306.17103 [cs].

[42] R. Yuan, H. Lin, Y. Wang, Z. Tian, S. Wu, T. Shen, G. Zhang, Y. Wu, C. Liu, Z. Zhou, Z. Ma, L. Xue, Z. Wang, Q. Liu, T. Zheng, Y. Li, Y. Ma, Y. Liang, X. Chi, R. Liu, Z. Wang, P. Li, J. Wu, C. Lin, Q. Liu, T. Jiang, W. Huang, W. Chen, E. Benetos, J. Fu, G. Xia, R. Dannenberg, W. Xue, S. Kang, and Y. Guo, "Chatmusician: Understanding and generating music intrinsically with llm," Feb. 2024. arXiv:2402.16153 [cs].

[43] 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," Jan. 2023. arXiv:2301.11325 [cs].

[44] P. Dhariwal, H. Jun, C. Payne, J. W. Kim, A. Radford, and I. Sutskever, "Jukebox: A generative model for music," Apr. 2020. arXiv:2005.00341 [eess].

[45] J. Copet, F. Kreuk, I. Gat, T. Remez, D. Kant, G. Synnaeve, Y. Adi, and A. Défossez, "Simple and controllable music generation,"

[46] S. Liu, A. S. Hussain, Q. Wu, C. Sun, and Y. Shan, "M^{2}ugen: Multi-modal music understanding and generation with the power of large language models," Dec. 2024. arXiv:2311.11255 [cs].

[47] S. Ding, Z. Liu, X. Dong, P. Zhang, R. Qian, J. Huang, C. He, D. Lin, and J. Wang, "Songcomposer: A large language model for lyric and melody generation in song composition," May 2025. arXiv:2402.17645 [cs].

[48] Y. Zhang, A. Maezawa, G. Xia, K. Yamamoto, and S. Dixon, "Loop copilot: Conducting ai ensembles for music generation and iterative editing," Aug. 2024. arXiv:2310.12404 [cs].

[49] 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. arXiv:2305.11000 [cs].

[50] P. K. Rubenstein, C. Asawaroengchai, D. D. Nguyen, A. Bapna, Z. Borsos, F. d. C. Quitry, P. Chen, D. E. Badawy, W. Han, E. Kharitonov, H. Muckenhirn, D. Padfield, J. Qin, D. Rozenberg, T. Sainath, J. Schalkwyk, M. Sharifi, M. T. Ramanovich, M. Tagliasacchi,

A. Tudor, M. Velimirović, D. Vincent, J. Yu, Y. Wang, V. Zayats, N. Zeghidour, Y. Zhang, Z. Zhang, L. Zilka, and C. Frank, "Audiopalm: A large language model that can speak and listen," June 2023. arXiv:2306.12925 [cs].

[51] J. Wu, Y. Gaur, Z. Chen, L. Zhou, Y. Zhu, T. Wang, J. Li, S. Liu, B. Ren, L. Liu, and Y. Wu, "On decoder-only architecture for speech-to-text and large language model integration," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, p. 1–8, Dec. 2023.

[52] S. Wang, C.-H. Yang, J. Wu, and C. Zhang, "Can whisper perform speech-based in-context learning?," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 13421–13425, Apr. 2024.

[53] W. Yu, C. Tang, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, Z. Ma, and C. Zhang, "Connecting speech encoder and large language model for asr," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 12637–12641, Apr. 2024.

[54] C. Wang, S. Chen, Y. Wu, Z. Zhang, L. Zhou, S. Liu, Z. Chen, Y. Liu, H. Wang, J. Li, L. He, S. Zhao, and F. Wei, "Neural codec language models are zero-shot text to speech synthesizers," Jan. 2023. arXiv:2301.02111 [cs].

[55] H. Hao, L. Zhou, S. Liu, J. Li, S. Hu, R. Wang, and F. Wei, "Boosting large language model for speech synthesis: An empirical study," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, p. 1–5, Apr. 2025.

[56] Z. Du, J. Wang, Q. Chen, Y. Chu, Z. Gao, Z. Li, K. Hu, X. Zhou, J. Xu, Z. Ma, W. Wang, S. Zheng, C. Zhou, Z. Yan, and S. Zhang, "Lauragpt: Listen, attend, understand, and regenerate audio with gpt," July 2024. arXiv:2310.04673 [cs].

[57] S. Kakouros, J. Šimko, M. Vainio, and A. Suni, "Investigating the utility of surprisal from large language models for speech synthesis prosody," June 2023. arXiv:2306.09814 [eess].

[58] Z. Liu, Z. Lai, Z. Gao, E. Cui, Z. Li, X. Zhu, L. Lu, Q. Chen, Y. Qiao, J. Dai, and W. Wang, "Controlllm: Augment language models with tools by searching on graphs," in *Computer Vision – ECCV 2024* (A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, eds.), (Cham), p. 89–105, Springer Nature Switzerland, 2025.

[59] C. Li, H. Chen, M. Yan, W. Shen, H. Xu, Z. Wu, Z. Zhang, W. Zhou, Y. Chen, C. Cheng, H. Shi, J. Zhang, F. Huang, and J. Zhou, "Modelscope-agent: Building your customizable agent system with open-source large language models," Sept. 2023. arXiv:2309.00986 [cs].