1 Abstract

Enhancing music with machine learning methods is challenging; it is essential to acknowledge the wisdom in existing approaches. The proposed solution to this wicked problem is a frequency mask that solves this problem by filtering the audio, similar to existing traditional approaches. With the use of multi-head attention, an auto-regressive loss function to incorporate both magnitude, temporal features, and spectral features, along with mapped normalized difference values to compete in the cadenza challenge[7, 6], it was not able to outperform whisper and performed comparable to traditional methods.

2 Introduction

Across multiple iterations, the following are the results of what was deemed adequate and what failed. An autoregressive token prediction model that utilizes greedy audio token prediction. This approach has proven to create an effective small model for roughly matching the root mean squared error of our digital footprint of spoken words, although the words generally sound different. However, it only functioned in applications that did not involve any other or multiple sounds at the same time. Models that take into account temporal and spectral audio features require significantly larger sizes to achieve satisfactory performance, requiring many more parameters than were available to achieve a good result. Whilst spectral mapping proved difficult to train the model at all. It is essential not to forget mathematics.

3 Design Decisions

Using a log magnitude spectrum and a short-time Fourier Transform, where the values had to be normalized to ensure that the features were learnable. With an 8-layer 2D convolutional U-Net architecture [8] made of 4 attention heads [2] , training across 32 batches, despite the small batch size of 2, allows enough generalization for learning. Using leaky ReLU and the AdamW optimizer [5, 9] with a normalization method that allows for negative values and skip connections [1, 4], which was crucial for a more accurate mathematical mapping of the data features. The loss function comprised weighted magnitude loss, log magnitude loss, and spectral convergence, which was chosen to balance energy across the speech frequencies[3]. **Explicitly, the method functions as a discriminative masker. It takes the Unprocessed Signal and predicts a mask to filter out noise. The loss function minimizes the difference between the masked output and the Clean Signal using a combination of Magnitude Loss and Spectral Convergence."**

3.1 Training environment

Training was done on a single NVIDIA RTX 3070 Ti (8GB VRAM), with a linear warm-up of 5 epochs, and was stopped after 30 epochs.

3.2 Post

Post-processing was performed using a Transient Booster, which involved calculating the derivative of the masked signal and then applying a gain to parts of the audio with a positive gradient.

4 Results

The model was not able to achieve higher results then the baseline. The performance overall was by this metric worse in both RMSE error and correlation.

System	**RMSE (Lower is better)**	**Correlation (Higher is better)**
Baseline Whisper	29.08	0.58
Your System (T050)	36.08	0.45
Result	error is higher (+7.00)	correlation is lower (-0.13)

4.1 Opinion

The filtered audio from the model sounds clearer then the unfiltered audio. Suspecting that this might be due to the filtering nature of the model has blurred the sounds making it harder for whisper to detect individual sounds.

5 Acknowledgments

All code in this project is generated by artificial intelligence, and it was well checked to the best of my ability. It may require additional oversight. The previous examples of the 2024 and 2023 Candanza challenge provided the initial inspiration for the validity of these solutions. The existing designs of Sound Stream, Eva AI, U-net, bit-net, Hifi-decoder, Haspi, and pyclarity, encodec, pytorch, torchaudio, "The author would like to acknowledge the assistance of Large Language Models in the debugging and architectural refinement of this project. Specifically,Gemini (Google) for architectural design and error resolution,Claude (Anthropic) for loss function optimization and peer review, and Qwen(Alibaba) for stability analysis. Special thanks to the concept of the 'Orion Train' for providing the morale to continue through hardware constraints."

References

[1] Zalán Borsos et al. *SoundStorm: Efficient Parallel Audio Generation*. May 16, 2023. DOI: 10.48550/arXiv.2305.09636. arXiv: 2305.09636[cs]. URL: http://arxiv.org/abs/2305.09636 (visited on 03/30/2025).

[2] "Fast Spectrogram Inversion Using Multi-Head Convolutional Neural Networks". en. In: *ResearchGate* (Aug. 2025). DOI: 10.1109/LSP.2018.2880284. URL: https://www.researchgate.net/publication/328843127_Fast_Spectrogram_Inversion_Using_Multi-Head_Convolutional_Neural_Networks (visited on 11/28/2025).

[3] Xilin Jiang et al. *AAD-LLM: Neural Attention-Driven Auditory Scene Understanding*. version: 1. Feb. 24, 2025. DOI: 10.48550/arXiv.2502.16794. arXiv: 2502.16794[cs]. URL: http://arxiv.org/abs/2502.16794 (visited on 03/30/2025).

[4] Zhifeng Kong et al. *DiffWave: A Versatile Diffusion Model for Audio Synthesis*. Mar. 30, 2021. DOI: 10.48550/arXiv.2009.09761. arXiv: 2009.09761[eess]. URL: http://arxiv.org/abs/2009.09761 (visited on 03/30/2025).

[5] Ilya Loshchilov and Frank Hutter. *Decoupled Weight Decay Regularization*. arXiv:1711.05101 [cs]. Jan. 2019. DOI: 10.48550/arXiv.1711.05101. URL: http://arxiv.org/abs/1711.05101 (visited on 11/28/2025).

[6] Gerardo Roa-Dabike et al. "The Cadenza Lyric Intelligibility Prediction (CLIP) Dataset". In: *Data in Brief* (2025).

[7] Gerardo Roa-Dabike et al. "Overview of the ICASSP 2026 Cadenza Challenge: Predicting Lyric Intelligibility". In: *Proc. IEEE ICASSP*. To appear. 2026.

[8] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. arXiv:1505.04597 [cs]. May 2015. DOI: 10.48550/arXiv.1505.04597. URL: http://arxiv.org/abs/1505.04597 (visited on 11/28/2025).

[9] Bing Xu et al. *Empirical Evaluation of Rectified Activations in Convolutional Network*. arXiv:1505.00853 [cs]. Nov. 2015. DOI: 10.48550/arXiv.1505.00853. URL: http://arxiv.org/abs/1505.00853 (visited on 11/28/2025).

6 Appendix

7 Model examination

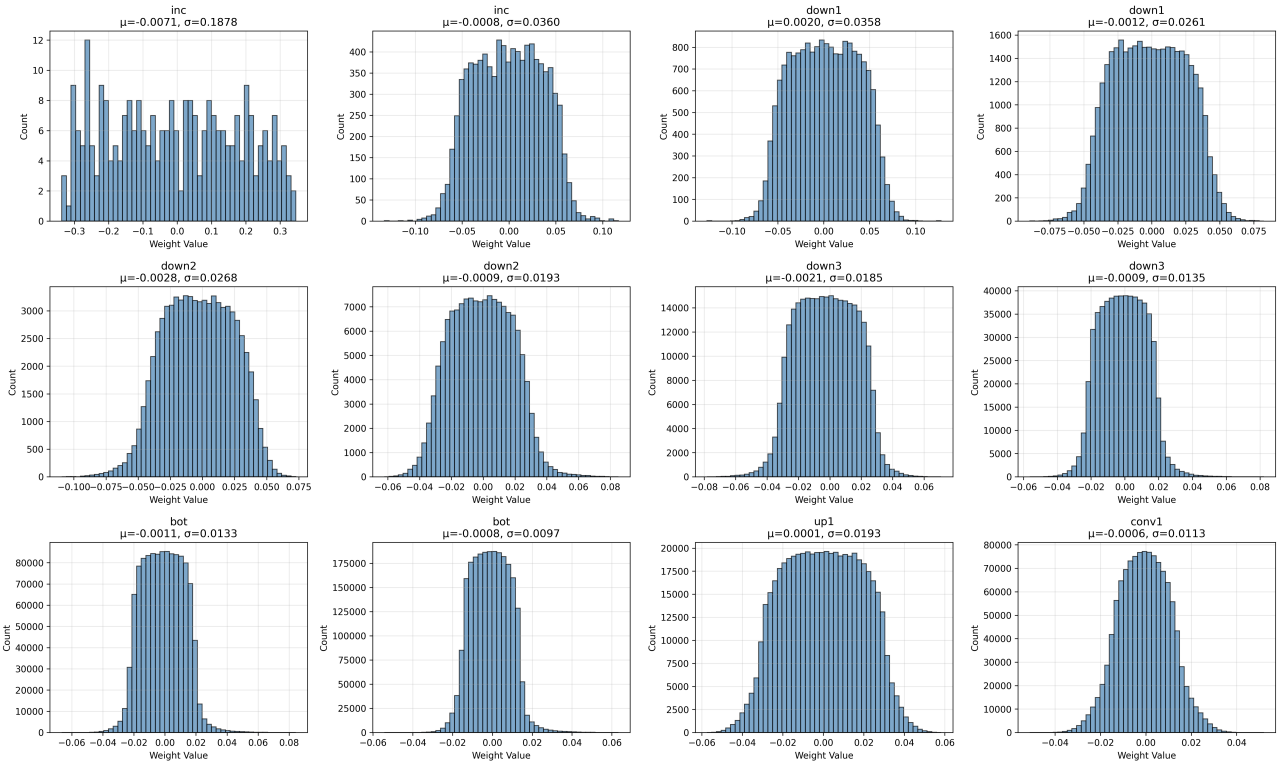

Figure 1: This shows how values progressed through the model.

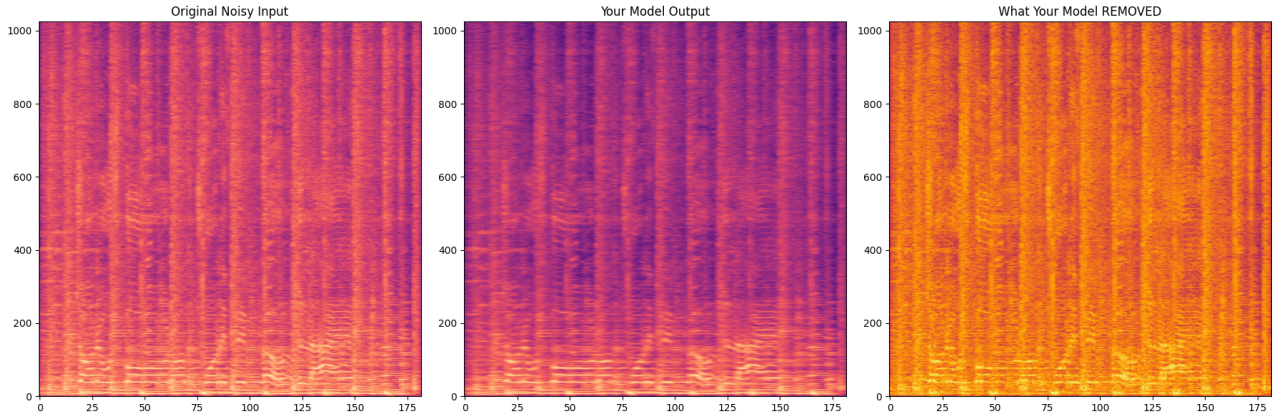

Figure 2: Shows how the model removed unwanted noise.

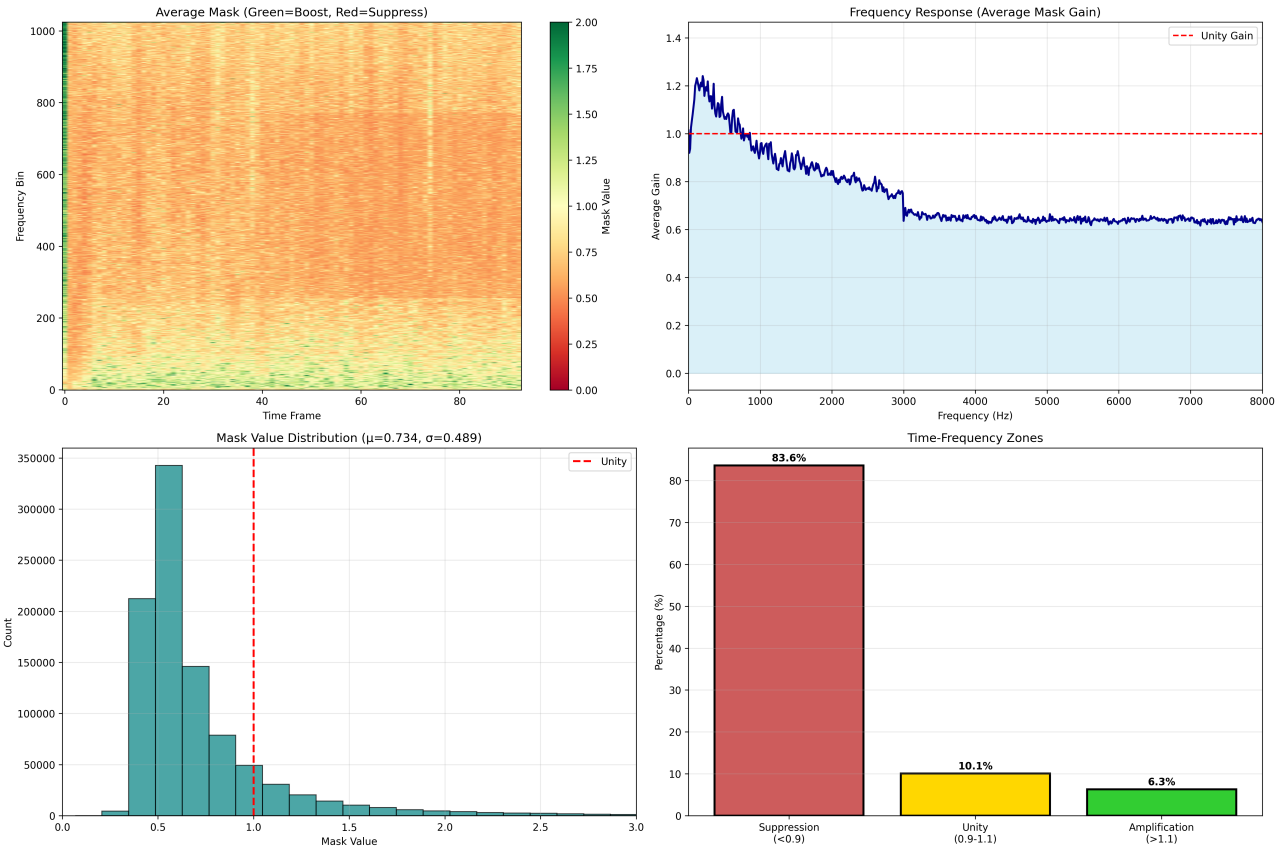

Figure 3: Shows areas target by masking of the model.

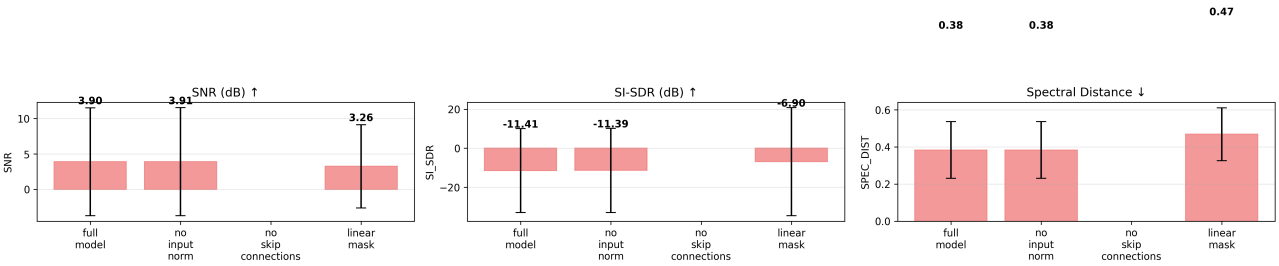

Figure 4: Ablation Results showing Signal-to-Noise Ratio across different model configurations.

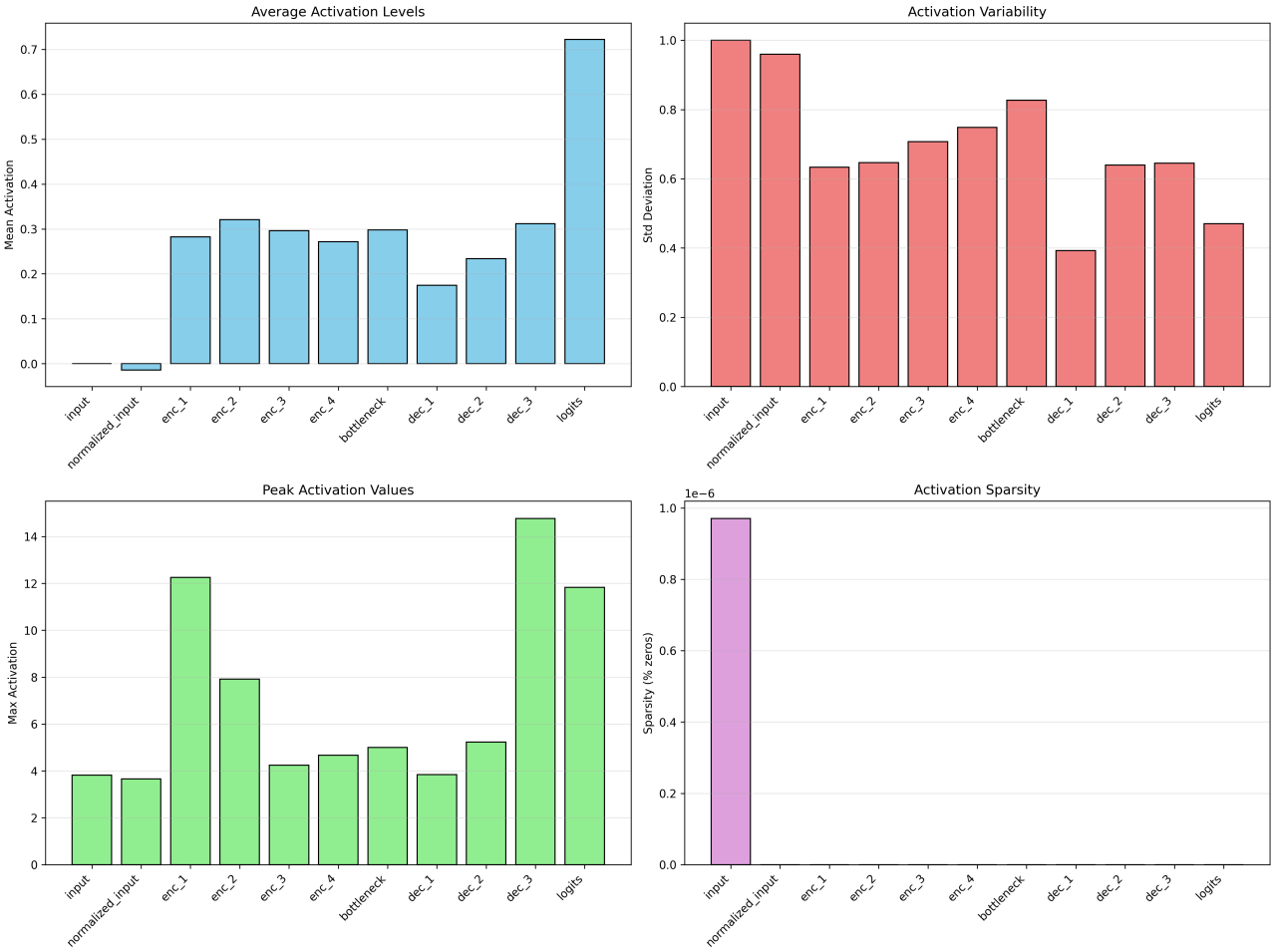

Figure 5: Demonstrates that the whole model was used with an activation pattern.

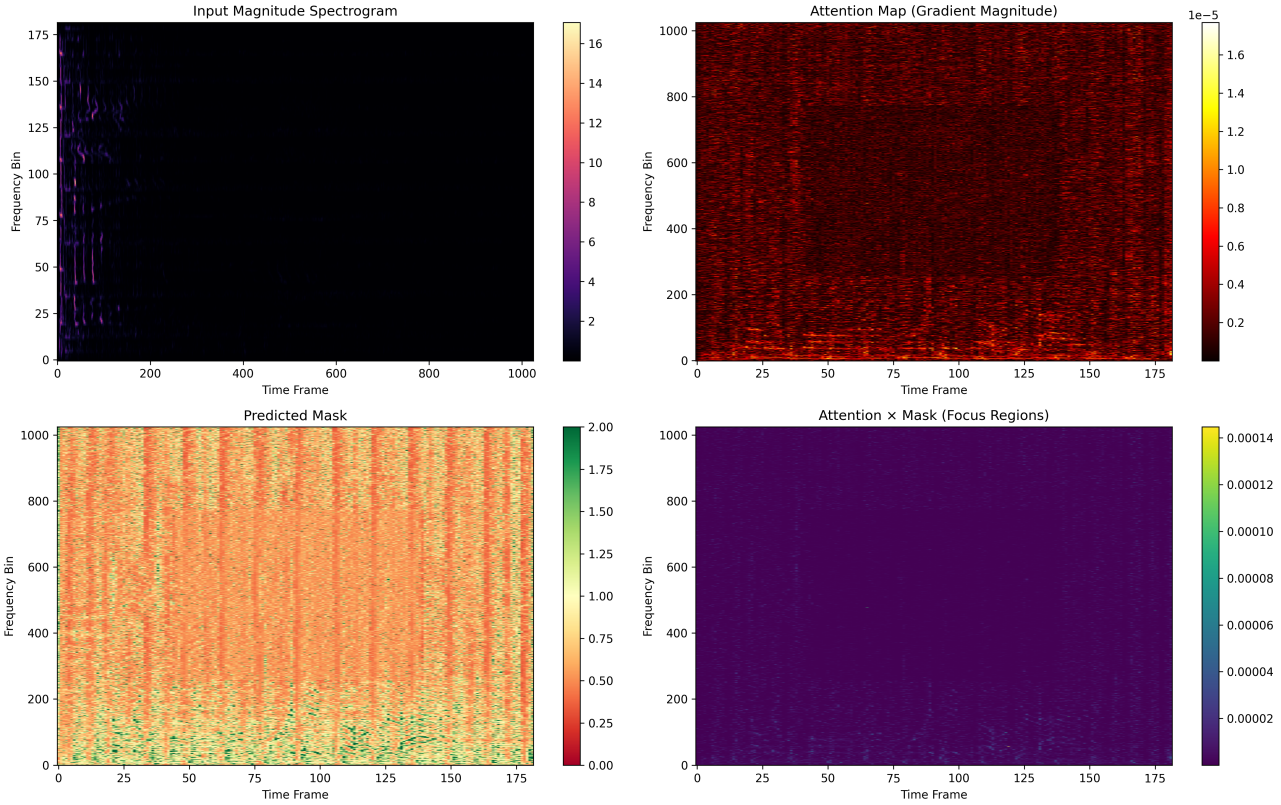

Figure 6: Attention maps

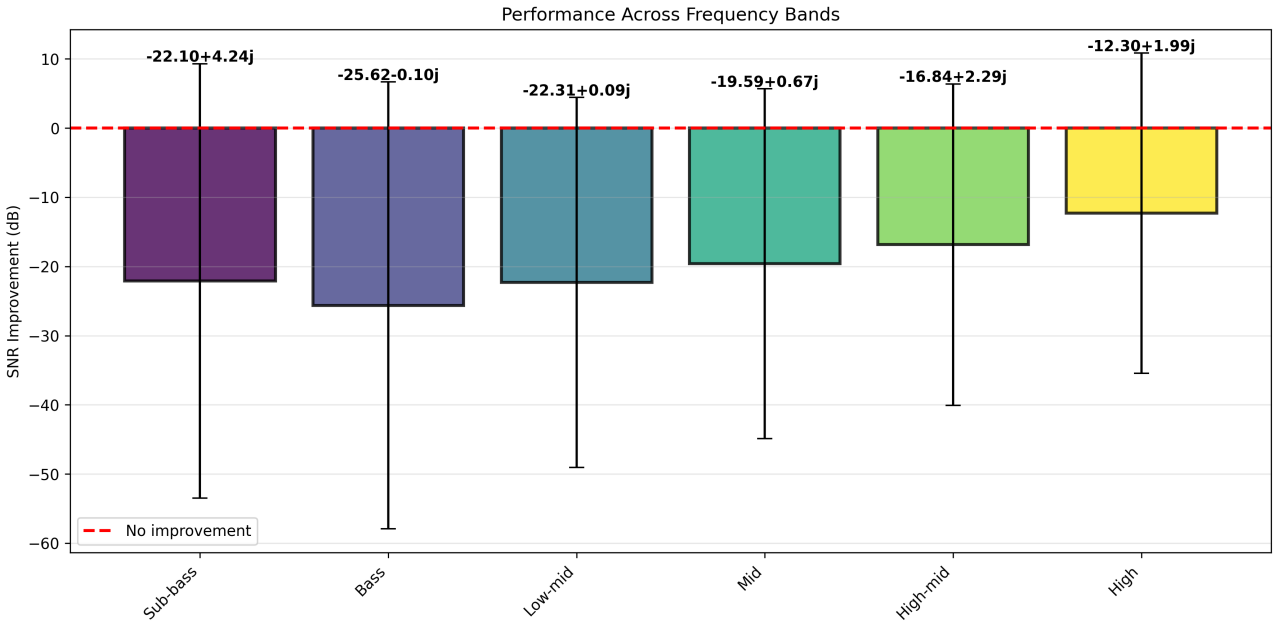

Figure 7: SNR Improvement per Frequency Band.

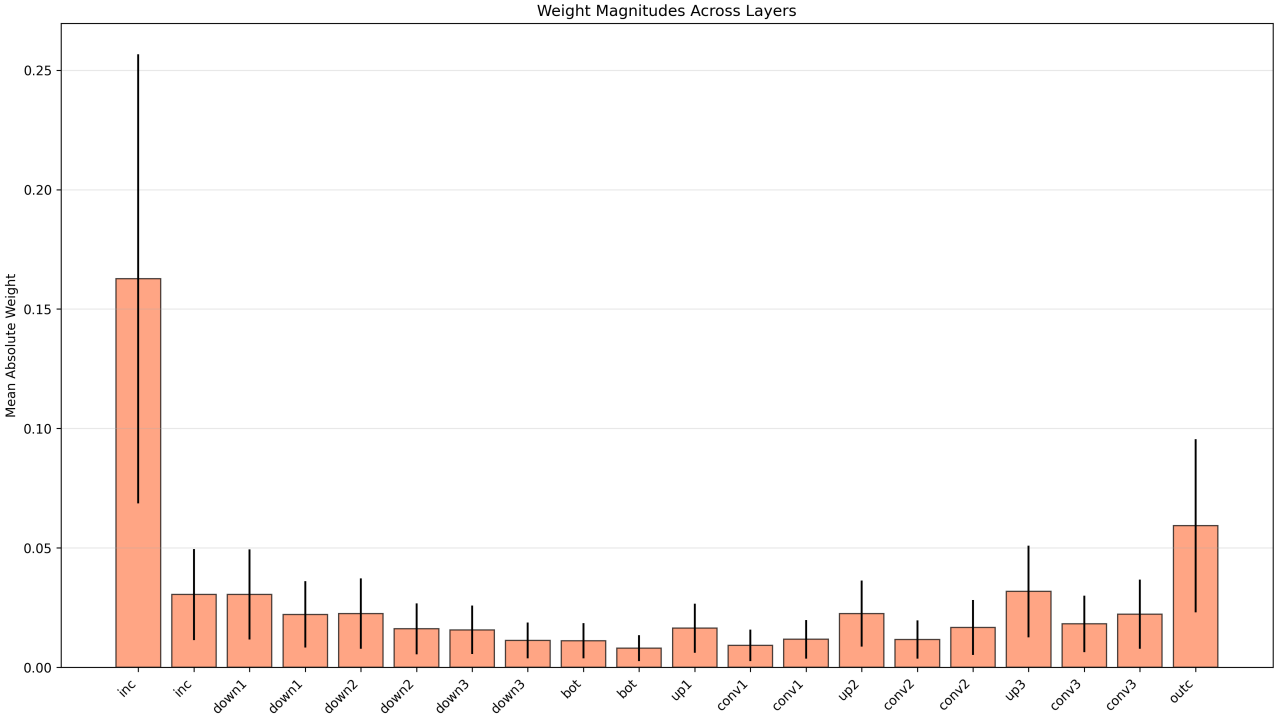

Figure 8: weight magnitudes across the model