Language-to-Sound Transformation Model using Archetype-Based Audio Representation (LSTMABAR)

By Shanth Gopalswamy, Zain Nasim

Abstract

Audio production suffers a critical gap: humans describe sound semantically, yet creation requires precise, technical digital signal processing (DSP) parameters. We introduce LSTMABAR (Language-to-Sound Transformer Model using Archetype-Based Audio Representation), a novel multimodal framework designed to bridge this divide with an emphasis on interpretability and controllability. LSTMABAR employs a two-tower architecture, using Sentence-BERT and ResNet-18 and is trained on the MusicCaps dataset via InfoNCE contrastive loss to align natural language with acoustic features. The system uniquely maps the resulting joint embedding to an interpretable mixture of five fundamental waveform archetypes (Sine, Square, Sawtooth, Triangle, Noise), which in turn drive a Differentiable Digital Signal Processing (DDSP) transformation engine. Crucially, the model incorporates a quantum-enhanced attention mechanism to capture complex, non-linear text-audio relationships. Through experimentation and hyperparameter tuning, the best model configuration utilized the quantum attention mechanism (8-qubit, depth 3, high regularization), providing superior semantic-to-parameter mapping. LSTMABAR lays the groundwork for a practical and interpretable method for language-driven manipulation of existing audio that can be expanded upon for a user-ready sound design tool.

Introduction

Audio production and sound design remain highly technical domains, requiring extensive knowledge of digital signal processing (DSP) parameters such as filter cutoff frequencies, resonance values, attack/decay envelopes, and harmonic content ratios. A sound engineer trying to create a "warm, smooth piano tone with gentle sustain" has to translate this intuitive description into precise numerical adjustments across dozens of parameters in complex audio production software. That gap is the problem: humans describe audio semantically, while machines require precise DSP parameters. Musicians could benefit from a head start where they are no longer slowed down by parsing through sound libraries and patches before landing on a tone they are happy with.

Recent multimodal models like CLAP (audio+language) show that language can meaningfully align with sensory data in shared embedding spaces. They can *understand* what audio should sound like based on text, and vice-versa. But understanding alone isn't sufficient for audio production. Producers do not just want to find the right sound, they want to shape the sound they already have into something new. That means going beyond embedding similarity and actually manipulating acoustic properties in a controllable, interpretable way.

In this project, we introduce LSTMABAR (Language-to-Sound Transformation Model using Archetype-Based Audio Representation), a multimodal framework designed to bridge the gap between natural language descriptions and audio transformations through interpretable archetype representations. Our system breaks down audio into five fundamental waveform archetypes: sine, square, sawtooth, triangle, and noise. Each archetype corresponds to distinct perceptual and acoustic properties. A sine wave produces smooth, mellow tones focused on the fundamental frequency; square waves add aggressive, odd harmonics; sawtooth waves contribute bright, cutting character through full harmonic spectra; triangle waves offer softer, muted qualities; and noise provides textural, airy characteristics. By learning to map text descriptions to archetype mixture weights, our model provides an interpretable intermediate representation that explains *how* transformations occur, not just that they occur.

Our architecture employs a two-tower contrastive learning framework, adapting techniques from CLAP and MuLan to the audio transformation domain. A text encoder based on Sentence-BERT processes natural language descriptions, while a spectrogram-based audio encoder using ResNet-18 extracts acoustic features. These embeddings are aligned in a shared space via InfoNCE contrastive loss, maximizing similarity for matched text-audio pairs while minimizing similarity for mismatched pairs. On top of that, we add a learned archetype prediction head that maps joint text-audio embeddings to five-dimensional archetype weights, and a DDSP (Differentiable Digital Signal Processing) transformation engine that applies targeted filtering, harmonic synthesis, and noise injection based on predicted weights.

Related Work

The development of the LSTMABAR framework builds upon recent advancements in multimodal learning, audio representation, and quantum-enhanced deep learning. The core of our approach aligns with the Contrastive Language-Audio Pretraining paradigm introduced by Elizalde et al. (2022), which demonstrated that projecting audio and text into a joint embedding space enables robust zero-shot performance. This architecture was further validated by Huang et al. (2025) with MuLan, which links music audio to unconstrained natural language using contextual BERT encoders.

Robust feature extraction is essential for high-fidelity synthesis. MuQ was introduced in 2025 demonstrating that Mel Residual Vector Quantization (Mel-RVQ) significantly enhances stability and efficiency in self-supervised learning compared to random projection. While our system uses a differentiable DSP engine rather than neural codes, MuQ's insights into tokenization inform our encoder design. Furthermore, the frameworks around speaker similarity and emotion recognition proposed by An et al. (2024) for FunAudioLLM provide a template for assessing the perceptual accuracy of our generated audio archetypes.

A distinct feature of our model is the quantum-enhanced attention mechanism. While Coecke et al. (2020) laid the groundwork in establishing that linguistic grammatical structures are quantum-native, Tomal et al. (2025) demonstrated that a hybrid classical-quantum Transformer could capture complex semantic relationships more efficiently than standard dot-product attention, inspiring our use of quantum layers to disentangle non-linear descriptor-parameter relationships. Finally, our use case evaluation pipeline draws on Huang et al. (2024), whose iterative human sampling methods for characterizing conversational tone offer a geometric framework for evaluating subjective auditory qualities.

Dataset

For our dataset we utilized the MusicCaps dataset hosted on Kaggle, a large-scale collection of music clips paired with rich natural language descriptions. The dataset comprises 5,521 audio samples, each 10 seconds in duration at 44.1 kHz sample rate. Unlike previous datasets that provide only genre or instrument labels, MusicCaps includes detailed textual descriptions of each audio sample written by musicians. Example descriptions include "pop, tinny wide hi hats, mellow piano melody, high pitched female vocal melody, sustained pulsating synth lead" providing the semantic richness necessary for learning nuanced audio-text correspondences. We employed an algorithmic keyword-matching approach to derive archetype mixture weights from MusicCaps text descriptions based on keyword lexicons for each of our 5 archetypes, such as "'smooth', pure, mellow" for sine and "bright', sharp, cutting", for sawtooth, etc. After filtering out the datapoints that were inaccessible, we partitioned the remaining 5,288 samples from the dataset into training (70%, 3701 samples), validation (15%, 793 samples), and test (15%, 794 samples) sets. The test set was held out entirely during development and used only for final evaluation. Audio samples were processed by loading 2.0-second segments at 44.1 kHz using librosa, and then padded or trimmed to exactly 88,200 samples for consistent tensor shapes. After this they were encoded as log-mel spectrograms using the following parameters: 2048-sample FFT windows, 512-sample hop length, 128 mel bands, and normalized mel-scale transformation. Spectrograms were converted to log-scale via log(mel + 1e-9) to compress dynamic range.

Audio embeddings were L2-normalized for contrastive learning. To improve robustness, we applied audio augmentation techniques such as random gain adjustment, random temporal shift, and additive Gaussian noise during training.

Methods

To address the challenge of interpretable text-to-audio synthesis, we developed the LSTMABAR (Language-Supervised Timbre Manipulation via Archetype-Based Audio Remixing) framework. Our system utilizes a two-tower architecture that aligns semantic descriptions with acoustic features to drive a differentiable digital signal processing synthesizer.

A. Architecture Overview

The system architecture bridges natural language and digital signal processing through a shared latent semantic space and consists of three primary stages: first the dual-encoder backbone to encode text and audio into a unified high-dimensional vector space, second the archetype prediction head to regress joint embeddings into interpretable synthesizer control parameters, and finally the differentiable physics engine to synthesize audio from those control parameters.

B. Dual-Encoder Backbone

The Dual-Encoder employs a two-tower architecture designed to learn a joint embedding space where semantically similar text descriptions and audio samples live close together. The text tower utilizes a pre-trained transformer architecture (sentence-transformers/all-MiniLM-L6-v2) to capture semantic nuances into tonal descriptions. Within the text encoder, we integrated a quantum-enhanced attention mechanism to capture complex, non-linear relationships between linguistic descriptors and acoustic features. This hybrid classical-quantum approach was chosen to disentangle intricate semantic-parameter mappings more efficiently than standard dot-product attention, ensuring the model could accurately translate abstract adjectives into precise physical archetype weights. The pooled output is then projected via an MLP (Multilayer Perceptron) to a shared embedding dimension $d = 768$. The audio encoder processes log-mel spectrograms extracted from input waveforms. We leverage a ResNet-18 backbone modified for single-channel audio input. This network extracts high-level acoustic features which are similarly projected to dimension d. Finally we take inspiration from CLAP in employing contrastive alignment. We minimize InfoNCE loss between normalized text and audio embeddings within a batch, scaled by a learnable temperature parameter.

C. Archetype Prediction

Bridging the gap between the latent embedding space and audio synthesis requires translating abstract embeddings into concrete synthesis parameters. To solve this, we implemented an archetype predictor, a specialized regression head composed of shared dense layers followed by two parallel output heads. The predictor takes the concatenated embeddings and outputs a control vector governing both timbre and dynamics. The timbre head (five parameters) outputs a softmax-normalized mixture vector corresponding to five fundamental oscillator archetypes – Sine, Square, Sawtooth, Triangle, and Noise.

D. Differentiable Physics Engine

The core generative component is the dynamic physics engine, a differentiable synthesizer that, unlike neural vocoders (e.g.: WaveNet), uses explicit signal processing math. For a given time and predicted parameters, the engine performs oscillator synthesis, timbre mixing, spectral filtering, and temporal shaping. The output is a normalized audio waveform differentiable with respect to the input control parameters, allowing gradients to flow from the audio output back to the prediction head.

E. Use Case Evaluation

To observe alignment of the model's objective interpretations of sound with subjective human perception, we implemented a use case evaluation pipeline. During this interactive phase the user is presented with a text prompt, and before-and-after audio samples for the user to rate on a scale of one to five based on perceptual alignment. This allows us to analyze how well the text-audio interpretation results translate to sound manipulation and provides a foundation for a future RLHF (Reinforcement Learning from Human Feedback) pipeline. We can then optimize the predictor parameters to maximize rewards using MSE loss which would complete a hybrid approach with both the robustness of supervised contrastive learning and the perceptual nuance of human-in-the-loop optimization.

Training

A. Hyperparameter Experiments

The training process was designed to identify the balance between model expressivity and generalization capabilities through rigorous hyperparameter tuning. We conducted systematic evaluation of the following dimensions: model architecture (classical vs. quantum attention), network depth, regularization strength, embedding dimension, batch size, learning rate, number of qubits, and noise strength.

During experimentation, we struggled with finding the balance between overfitting and underfitting. At first, all model configurations converged to near-zero training losses, showed promising contrastive loss curves,but exhibited divergent validation losses (see Figure B). To combat this we ran experiments using gradual unfreezing, increased dropout rate, and increased weight decay. However, these experiments led to underfitting where both training and validation data improved at an incredibly slow rate. Ultimately the best performing model was the *quantum_8qubit_depth3_high_reg* model (see Figure D.) with a validation loss of 1.7026.

B. Architectural & Design Experiments

Beyond hyperparameter tuning, the development of LSTMABAR involved significant architectural experimentation to determine the optimal balance between expressivity and control. Before converging on our combination of Neural DDSP and Direct Archetype Prediction, we explored several other synthesis backbones. These alternative backbones showed promise, but required greater architectural adjustments to see measurable impact.

Deterministic Physics Engine

We replaced the neural filters with hard-coded mathematical oscillators that matched the 5 weights of the predictor. We generated a synthetic "warm start", pre-training the predictor on this data to teach it basic physics before use case evaluation or, in the future, RLHF.

9 Dimension Predictor and Physics Engine

We expanded the bottle neck to 9 parameters: 4 timbre weights + 5 dynamic controls (attack, decay, sustain, release, filter cutoff). This allowed the model to map more semantic concepts of time ("pluck", "hit") to physical envelopes (fast attack, low sustain).

Macro-Control Latents

Instead of predicting raw physics directly, the model predicts 32 abstract “Macro-Controls” (timbre, envelope, modulation latents). A small decoder network expands these into high-fidelity parameters (64 harmonics, noise bands). This decouples meaning (latents) from physics (decoder), allowing the engine to use complex additive synthesis (64 sine waves) without forcing management of 64 individual sliders.

Pure Quantum NLP using lambeq

We replaced the feature extractor in the text tower with lambeq which converts natural language description to distributional compositional categorical (DisCoCat) diagrams and then into parameterized quantum circuits. Rather than relying on statistical attention, it treats language as a tensor network based on grammatical structure and captures word interaction via quantum interference. After experimentation, we opted against using this architecture because of high computational cost and low applicability as we have relatively little to gain from capturing word interactions.

Results and Discussion

The final *quantum_8qubit_depth3_high_reg* model was evaluated on the held-out test set of 794 samples. The model achieved a Mean Square Error (MSE) of 0.0639 and a Cosine Similarity of 0.7474 between generated and target audio embeddings (see Figure E). The high cosine similarity indicates that while the generated audio may not be a perfect waveform replica of the ground truth, it is semantically highly aligned with the target descriptions. The Pearson correlation of 0.5225 further suggests a moderate-to-strong linear relationship between the predicted synthesis parameters and the ground truth characteristics.

Breaking down performance by waveform archetype reveals distinct behaviors in how the model maps language to sound (see Figure I). Noise was the most accurately predicted archetype (lowest MSE: 0.0421, highest Correlation: 0.678). This suggests that textural descriptors for noise (e.g., “hiss”, “airy”, “static”) are semantically distinct and easily mapped to the noise oscillator. Conversely, sine waves proved the most difficult to predict (MSE: 0.11311). This is likely due to the semantic ambiguity of descriptors like “mellow” or “smooth,” which can overlap significantly with triangle waves. This confusion is reflected in the classification report (see Figure H), where triangle waves had the lowest F1-score (0.2406), frequently being misclassified as sine due to similar spectral properties.

A critical finding from our experimentation is the tradeoff between interpretability and expressivity. While a pure end-to-end DDSP approach (conditioning the synthesizer directly on embeddings) might maximize acoustic fidelity, it sacrifices interpretability. By forcing the model to predict explicit archetype weights, we maintained the ability to open the black box and understand why a sound was generated. However, this led to gaps on the generation end. While the training and test results suggested strong model performance, those results did not translate to use case evaluation. Our approach of producing audio from predicted archetype weights proved too constrictive, causing many produced audio samples to sound similarly convoluted. When offering a “warm start” to the use case evaluation phase, results improved measurably but were then overfitting and overreliant on the “warm start” scenarios.

We hypothesize that a more robust implementation of the macro-control latents architecture with data that is better tailored to our use case would lead to optimal results on both the language-audio interpretation and generation fronts. Both interventions would also potentially address the overfitting problem encountered in training experiments as it would allow the model to ignore noisy, irrelevant musical concepts found in the current dataset and force the model to learn generalizable semantic concepts rather than precise rigid oscillator weights. We intend to explore this design consideration in greater detail as an immediate next step.

Conclusion

In this work, we introduced LSTMABAR, a novel framework bridging the gap between semantic language descriptions and technical audio production. By leveraging a quantum-enhanced attention mechanism and a deterministic archetype-based representation, we demonstrated that it is possible to translate abstract natural language into precise, interpretable DSP parameters.

Our results show that the model captures complex non-linear relationships between text and audio, achieving high semantic alignment and robust archetype classification. While challenges remain in distinguishing spectrally similar waveforms and aligning musical datasets with sound design tasks, the system successfully moves beyond simple sound retrieval to controllable sound transformation. These successes did not come without failures, specifically as it pertains to audio generation in the use case evaluation phase. Future iterations will focus on refining the dataset to better align with timbral manipulation and expand on our experimentation with macro-control latents, ultimately empowering creators to sculpt sound using the natural language they already speak.

References

Anonymous, et al. (2025). "CLAMP3: Contrastive Language-Audio-Music Pre-training." https://arxiv.org/abs/2502.10362

Anonymous, et al. (2025). "MuQ: Self-Supervised Music Representation Learning." https://arxiv.org/abs/2501.01108

Tomal, S.M.Y.I., et al. (2025). "Quantum-Enhanced Attention Mechanism in NLP: A Hybrid Classical-Quantum Approach." https://arxiv.org/abs/2501.15630

An, K., Chen, Q., Deng, C., Du, Z., Gao, C., Gao, Z., Gu, Y., He, T., Hu, H., Hu, K., Ji, S., Li, Y., Li, Z., Lu, H., Luo, H., Lv, X., Ma, B., Ma, Z., Ni, C., Song, C., Shi, J., Shi, X., Wang, H., Wang, W., Wang, Y., Xiao, Z., Yan, Z., Yang, Y., Zhang, B., Zhang, Q., Zhang, S., Zhao, N., & Zheng, S. (2024). “FunAudioLLM: Voice understanding and generation foundation models for natural interaction between humans and LLMs.” https://arxiv.org/abs/2407.04051

Huang, D.-M., Van Rijn, P., Sucholutsky, I., Marjieh, R., & Jacoby, N. (2024). “Characterizing similarities and divergences in conversational tones in humans and LLMs by sampling with people.” https://arxiv.org/abs/2406.04278

Elizalde, B., et al. (2022). "CLAP: Learning Audio Concepts from Natural Language Supervision." https://arxiv.org/abs/2206.04769

Contributions

Shanth Gopalswamy: project ideation, project outline, literature review, architecture design, text_tower, audio_tower, contrastive_alignment, ddsp_transformation, archetype_predictor, musiccaps_loader, model training and experimentation

Zain Nasim: project ideation, project outline, literature review, architecture design, text_tower fine tuning, model training and experimentation, rlhf pipeline, model evaluation pipeline

Appendix

Figure A. Validation Loss by Configuration

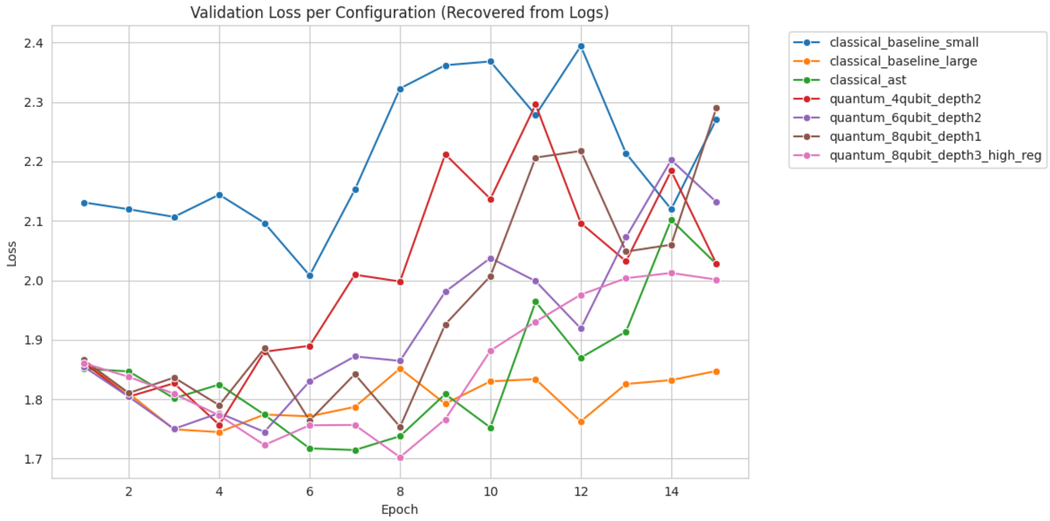


Figure B. Training/Validation Loss Curves

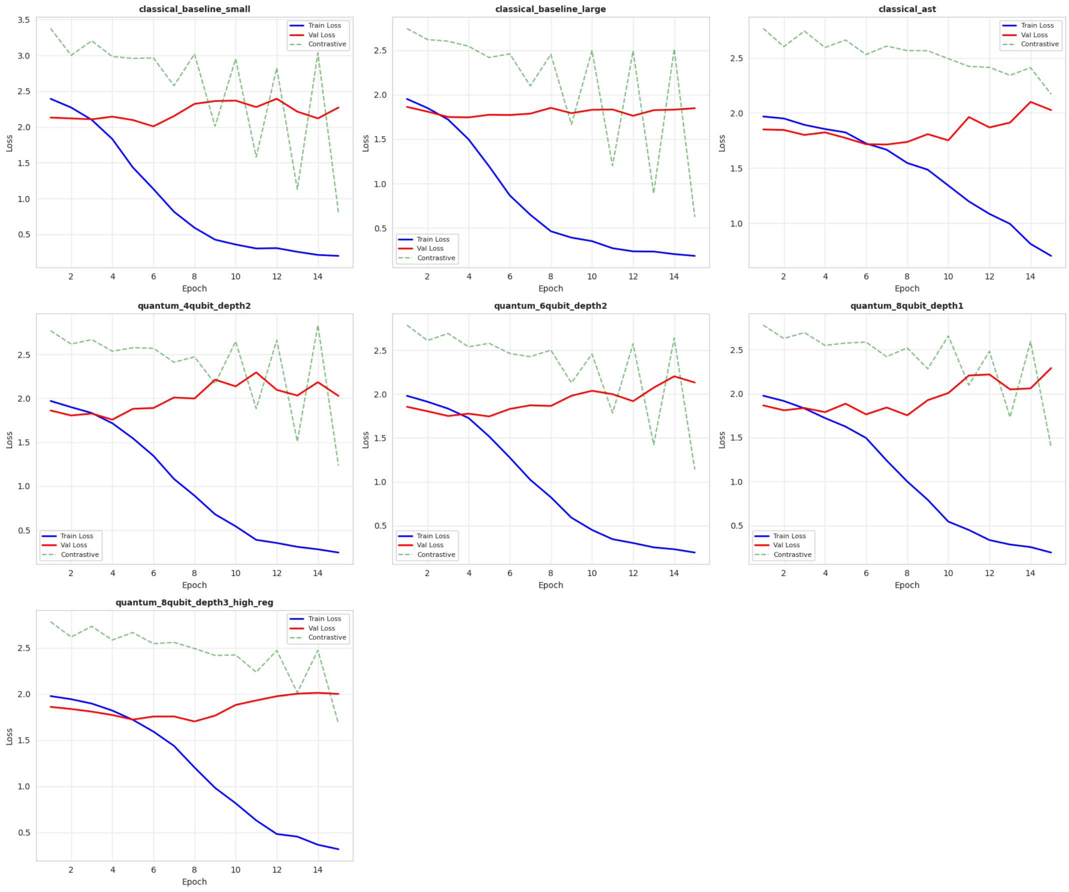


Figure C. Training Model Performance and Efficiency Plots

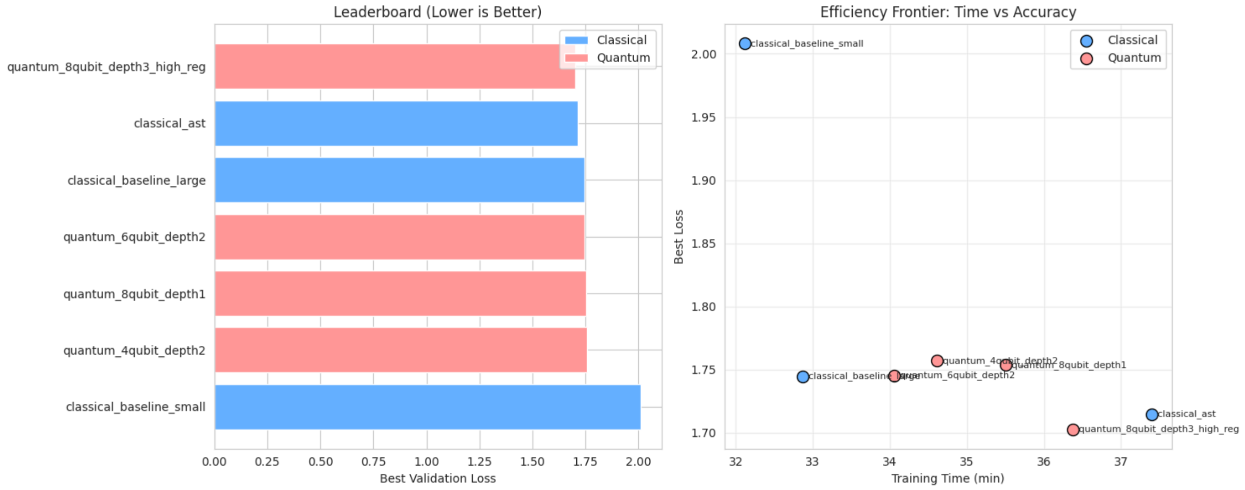
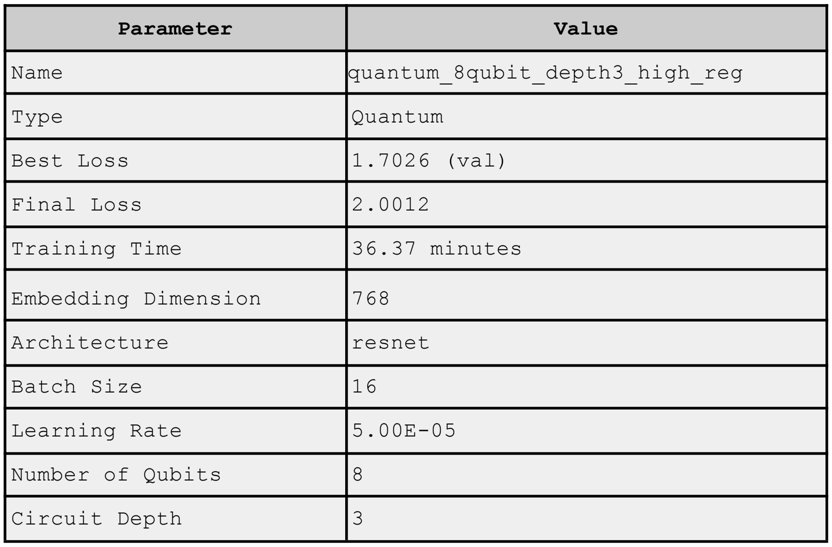
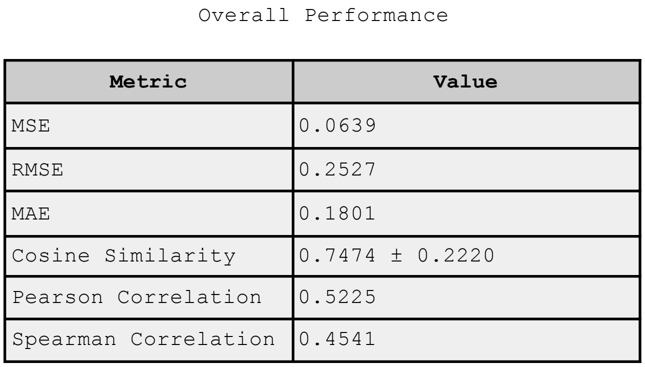


Figure D. Best Model Configuration



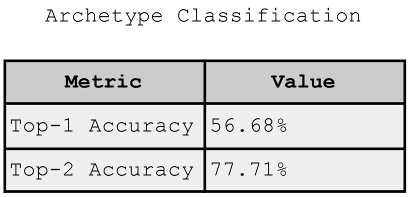
Parameter	Value
Name	quantum_8qubit_depth3_high_reg
Type	Quantum
Best Loss	1.7026 (val)
Final Loss	2.0012
Training Time	36.37 minutes
Embedding Dimension	768
Architecture	resnet
Batch Size	16
Learning Rate	5.00E-05
Number of Qubits	8
Circuit Depth	3

Figure E. Test Set Metrics



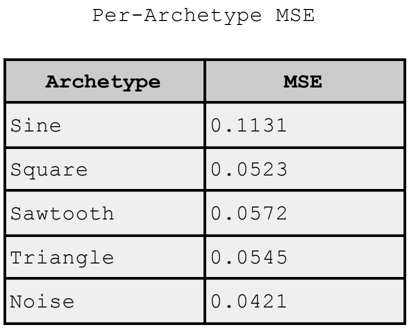
Overall Performance

Metric	Value
MSE	0.0639
RMSE	0.2527
MAE	0.1801
Cosine Similarity	0.7474 ± 0.2220
Pearson Correlation	0.5225
Spearman Correlation	0.4541



Archetype Classification

Metric	Value
Top-1 Accuracy	56.68%
Top-2 Accuracy	77.71%



Per-Archetype MSE

Archetype	MSE
Sine	0.1131
Square	0.0523
Sawtooth	0.0572
Triangle	0.0545
Noise	0.0421

Figure F. Predicted vs. Target Archetype Weights

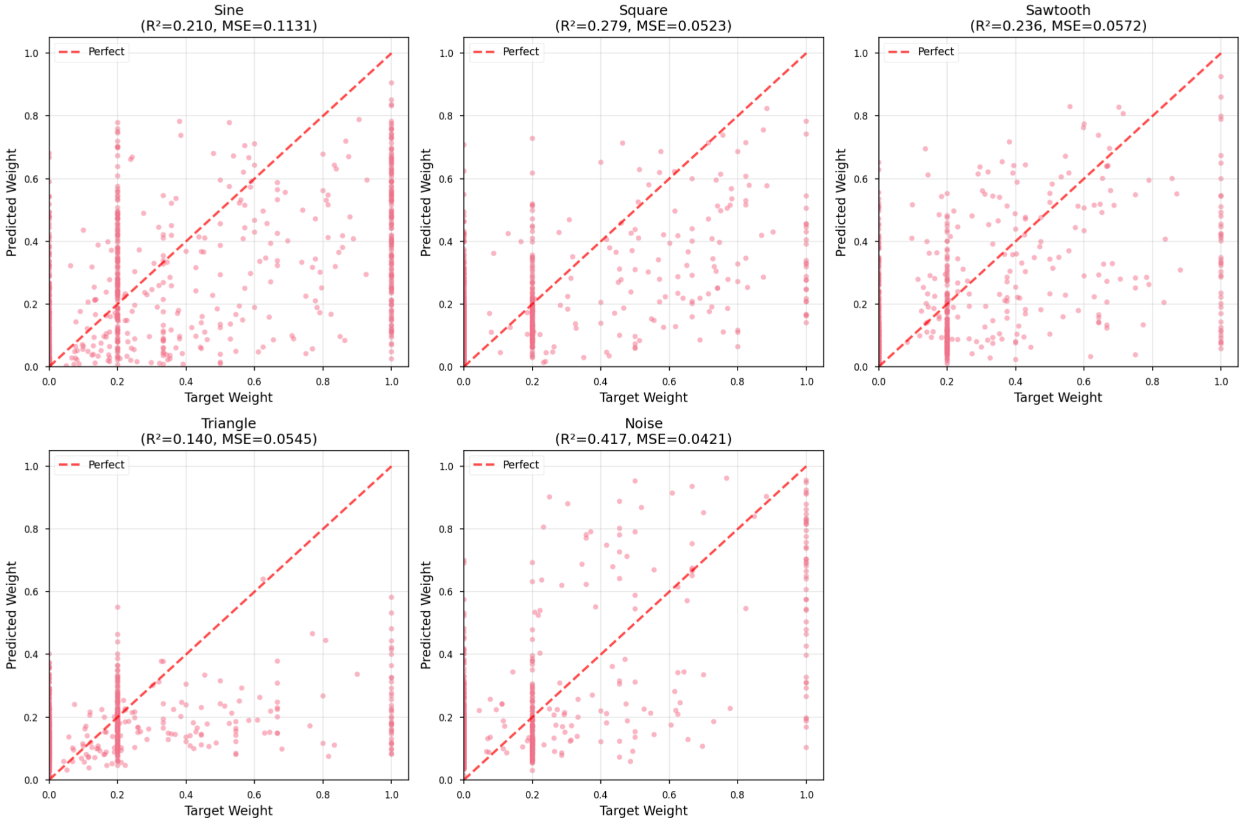


Figure G. Prediction Error

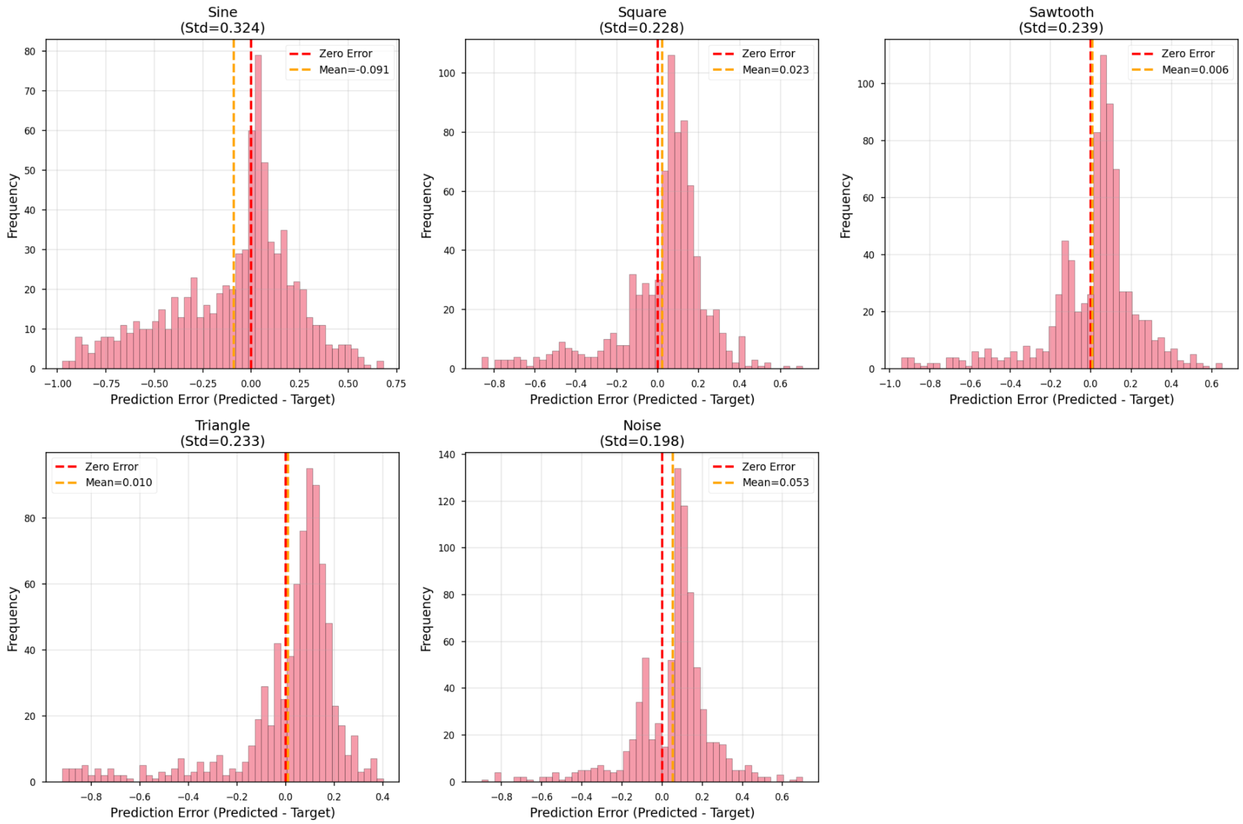
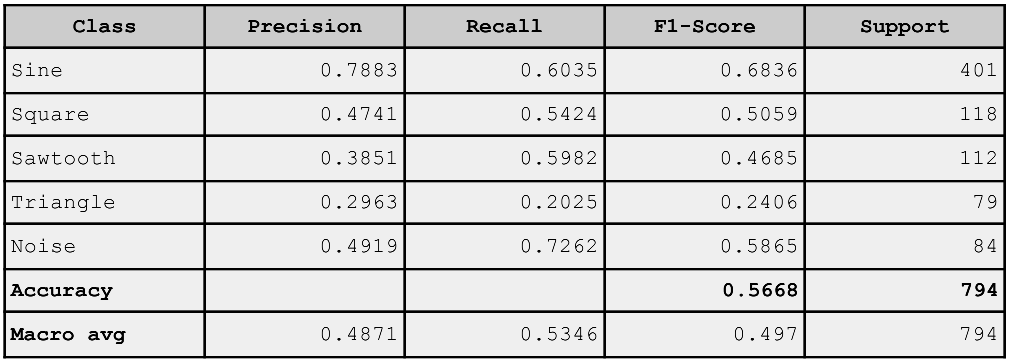
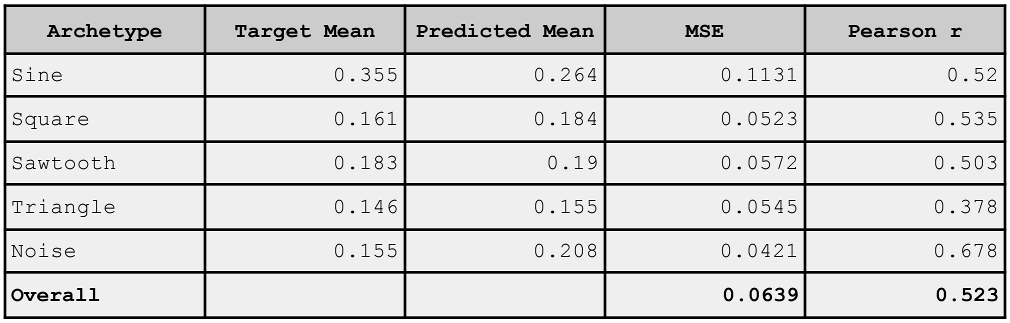


Figure H. Classification Report



Class	Precision	Recall	F1-Score	Support
Sine	0.7883	0.6035	0.6836	401
Square	0.4741	0.5424	0.5059	118
Sawtooth	0.3851	0.5982	0.4685	112
Triangle	0.2963	0.2025	0.2406	79
Noise	0.4919	0.7262	0.5865	84
Accuracy			**0.5668**	**794**
Macro avg	0.4871	0.5346	0.497	794

Figure I. Performance by Archetype



Archetype	Target Mean	Predicted Mean	MSE	Pearson r
Sine	0.355	0.264	0.1131	0.52
Square	0.161	0.184	0.0523	0.535
Sawtooth	0.183	0.19	0.0572	0.503
Triangle	0.146	0.155	0.0545	0.378
Noise	0.155	0.208	0.0421	0.678
Overall			**0.0639**	**0.523**