Quantizing Audio Diffusion Models

Nithya Shikarpur, Mason Wang, Asmi Kumar, Claire Lu

Contributions
Poster: Mason, Asmi, Nithya
Experiments/Evaluations: Mason, Nithya, Asmi
Visualizations: Mason, Asmi, Nithya
Report: Claire, Nithya, Asmi, Mason

INTRODUCTION

Challenges in Quantizing Diffusion Models

Diffusion models are challenging to quantize:

1. **Quantization errors accumulate** across diffusion steps.
2. **Activation Distributions vary** across diffusion steps.
3. **Skip Connections** in **decoder layers** result in **bimodal weight distributions** in the weights of the convolutional layers they are fed to.

Contribution 1: Quantizing Audio Diffusion Models

The work done in **Q-Diffusion** addresses many of the above challenges for **diffusion models** in the **image domain.** In this project, we:

1. Observe **corresponding trends** in **audio diffusion models.**
2. **Apply** different methods of **weight quantization** to an audio diffusion model for singing voice synthesis in Hindustani classical music (GaMaDHaNi).
3. **Evaluate** these methods on **audio-based** metrics.

Contribution 2: Adapting Quantization for Conditional Diffusion Models

While Q-Diffusion quantizes **unconditional** diffusion models, our project quantizes **conditional** diffusion models. Thus, we:

1. Find that for our **conditional diffusion model**, the activation magnitudes of **conditioning signals** also result in **multi-modal weight distributions** within both the **encoder and decoder layers.**
2. Apply a **per-group** method of quantizing weights to address the effect of **conditioning signal** and **skip connections simultaneously,** in both the **encoder** and **decoder**.

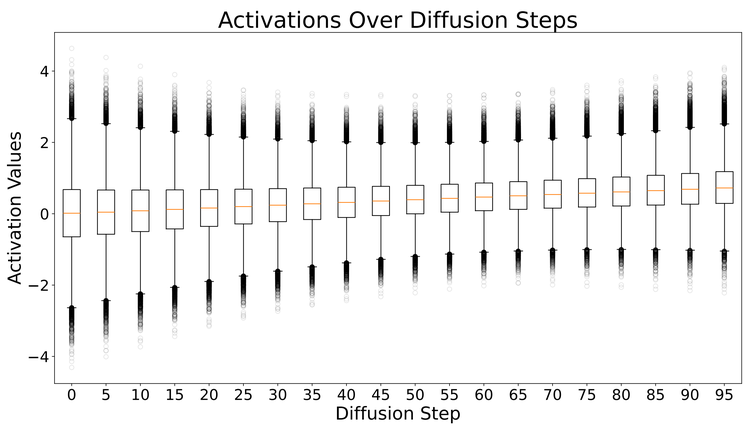


Figure 1 (Top): Activation magnitudes vary across diffusion steps. Unlike in Q-Diffusion, the progression is not monotonic.

Figure 3 (RIght): For the 8-bit per-tensor-quantized model, quantization errors accumulate, leading to divergence in step 50. In Base Model and our method, activations do not diverge.

INSIGHTS AND OBSERVATIONS

Each **decoder block** in the denoising network takes in the **previous layer's output**, the **skip connection**, and **three conditioning signals** (the time-step, singer ID and pitch contour condition). These are **concatenated along the channel dimension.**

As observed in Q-Diffusion, the skip connection has a much larger range of magnitudes than the output from the previous layer. **However**, the **conditioning signals** also have different ranges than either of them. (Fig. 3, Left)

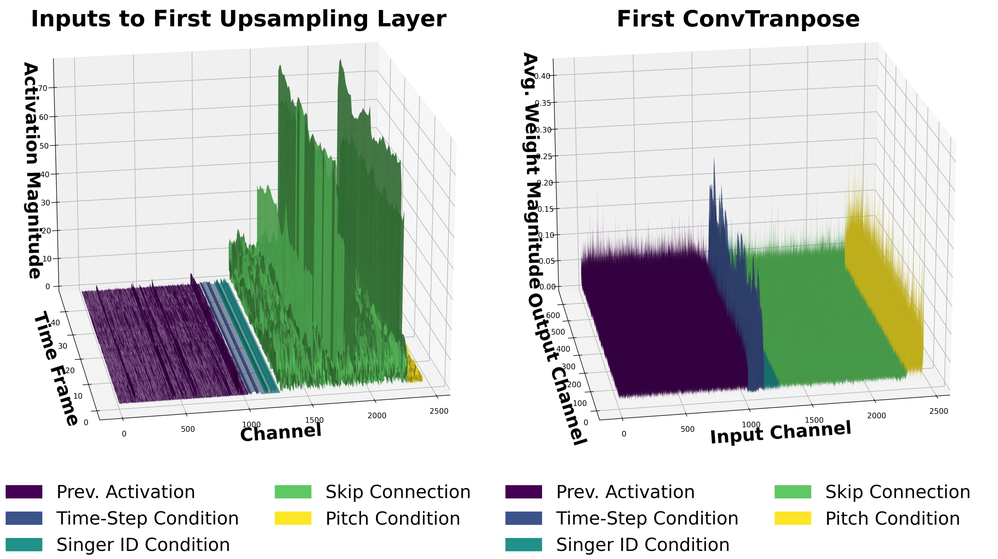


Figure 2.

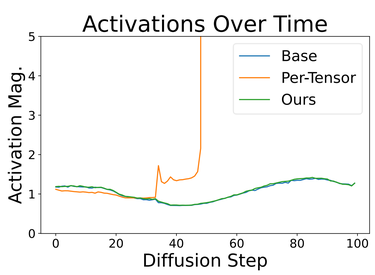
The first convolutional layer(s) in each decoder block learns weights with magnitudes **inversely proportional** to the magnitude of the input group (Right). This is true not only for the skip connections, but also for the conditioning signals.

This results in **inconsistent weight magnitudes** across the input channel dimension, making the weight tensor difficult to quantize.

METHOD

1. **Group input channels together** based on which part of the concatenated input they correspond to (Fig. 2, Right).
2. For all layers that take a concatenated input, **quantization statistics** (min, max) are computed separately per group of input channels.
3. Weights are quantized **per-output-channel**, and **per-input-channel group.**

Results



	FAD ↓	MCD ↓	LSD ↓
Base Model	214.7	694.0	21.2
4-Bit Per-Tensor	**394.4**	896.9	**38.7**
4-Bit Per-Channel	439.3	892.5	38.9
4-Bit Grouped (Ours)	415.0	**886.0**	39.2
8-Bit Per-Tensor	339.3	1015.6	37.5
8-Bit Per-Channel	220.1	693.7	**21.2**
8-Bit Grouped (Ours)	**214.5**	**687.5**	21.4

Table 1: Quantitative evaluations of our generated audio

EVALUATION

We assess performance of each quantization method against **16 high-quality ground-truth audio examples.** We extract the pitch contour from each ground-truth audio sample and use it to condition the generation of each quantized model.

Fréchet Audio Distance (FAD)

FAD captures the perceptual similarity between ground-truth and generated audio by comparing the distribution of OpenL3 embeddings, which represent high-level audio features such as timbre and tonality. Lower FAD indicates higher similarity.

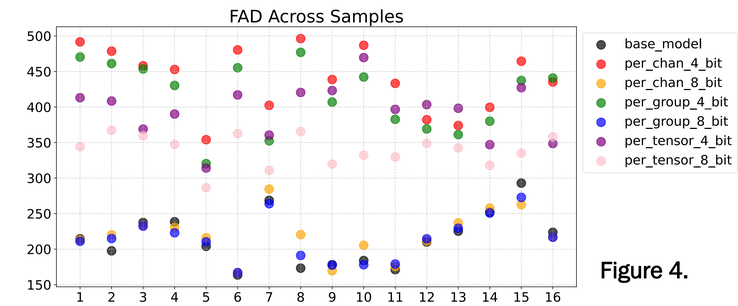


Figure 4.

The dot plot (Fig. 4) **preserves per-sample scores,** helping identify relationships between quantization methods on a per-sample basis. We **aggregate** the data in (Fig. 5)'s boxplot.

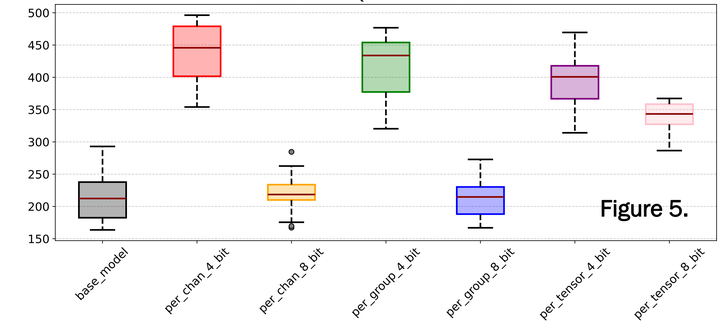


Figure 5.

Mel Cepstral Distortion (MCD)

MCD quantifies the spectral distortion between two audio signals, comparing Mel-Frequency Cepstral Coefficients aligned using Dynamic Time Warping. Lower MCD scores indicate more preservation of timbral and pitch characteristics. *Essentially identical trends in MCD are shown at tinyurl.com/mcd-metric.*

Log Spectral Distance (LSD)

LSD measures the distortion in the frequency domain by comparing the log-scaled power spectra of the audio. Lower LSD scores indicate better preservation of spectral fidelity. *Essentially identical trends in LSD are shown at tinyurl.com/lsd-metric.*

DISCUSSION

8-bit per-channel and 8-bit grouped quantization significantly outperform 4-bit methods and perform comparably to the base model across all metrics in all samples. 8-bit methods also exhibit stronger clustering and stability across all 16 samples, underscoring their robustness. These results suggest that **channel-level and group-level 8-bit quantization** are practical solutions for deploying high-quality audio models efficiently.