ROBUST RESIDUAL FINITE SCALAR QUANTIZATION FOR NEURAL COMPRESSION

Xiaoxu Zhu, Jiakui Li, Ken Zheng, Guiping Zhong, Huimeng Wang, Shiyin Kang⋆*, Dahua Lin*

SenseTime Research, Beijing, China

ABSTRACT

Finite Scalar Quantization (FSQ) offers simplified training but suffers from residual magnitude decay in multi-stage settings, where subsequent stages receive exponentially weaker signals. We propose Robust Residual Finite Scalar Quantization (RFSQ), addressing this fundamental limitation through two novel conditioning strategies: learnable scaling factors and invertible layer normalization. Our experiments across audio and image modalities demonstrate RFSQ's effectiveness and generalizability. In audio reconstruction at 24 bits/frame, RFSQ-LayerNorm achieves 3.646 DNSMOS, a 3.6% improvement over state-of-the-art RVQ (3.518). On ImageNet, RFSQ achieves 0.102 L1 loss and 0.100 perceptual loss, with LayerNorm providing 9.7% L1 improvement and 17.4% perceptual improvement over unconditioned variants. The LayerNorm strategy consistently outperforms alternatives by maintaining normalized input statistics across stages, effectively preventing exponential magnitude decay that limits naive residual approaches. RFSQ combines FSQ's simplicity with multi-stage quantization's representational power, establishing a new standard for neural compression across diverse modalities.

***Index Terms*—** Neural compression, finite scalar quantization, residual quantization, audio coding, image compression

1. INTRODUCTION

Vector quantization has been a cornerstone of neural compression since the foundational work of Gray [1], establishing the theoretical framework for discrete representation learning in neural networks. The introduction of VQ-VAE [2] marked a paradigm shift by revolutionizing discrete representation learning, enabling end-to-end training of quantized neural networks while maintaining differentiability through straight-through estimators. Despite its groundbreaking contributions, traditional VQ methods encounter persistent challenges including codebook collapse phenomena, training instability issues, and the necessity for carefully tuned auxiliary losses to maintain codebook utilization [3].

Finite Scalar Quantization (FSQ) [4] simplifies quantization by independently quantizing each dimension to predetermined values, eliminating learnable codebooks while maintaining high-quality reconstruction. FSQ has shown success in speech synthesis [5, 6], speech recognition [7], and low-bitrate coding [8].

Limitations and Our Approach: Despite these successes, FSQ faces a fundamental limitation: its fixed, axis-aligned quantization boundaries may not optimally capture complex data distributions. While residual quantization could theoretically address this through progressive refinement, naive application suffers from the **residual magnitude decay problem**, where subsequent stages receive progressively weaker signals, severely limiting their effectiveness.

Our Contributions.

1. We identify and analyze the residual magnitude decay problem in naive residual FSQ implementations, providing theoretical and empirical evidence of its impact.

2. We propose two novel conditioning strategies—learnable scaling and invertible LayerNorm—that robustly address this problem while maintaining FSQ's simplicity.

3. We conduct comprehensive experiments on both audio and image reconstruction tasks, demonstrating significant improvements over strong baselines.

4. We establish RFSQ as a general framework applicable to various architectures and modalities, with consistent performance gains.

2. RELATED WORK

Vector quantization [2, 3] established discrete representation learning but suffers from codebook collapse and training instability [9]. FSQ [4] eliminates codebook learning by quantizing dimensions independently to predefined levels, showing success in image generation [10] and speech synthesis [5, 6].

Residual VQ [11, 12] applies hierarchical quantization for fine-grained detail capture. EnCodec [12] demonstrates RVQ's effectiveness for audio compression. However, naive residual FSQ suffers from magnitude decay, limiting subsequent stages' effectiveness.

⋆Corresponding author

3. METHOD

3.1. Background: Finite Scalar Quantization

FSQ quantizes a d-dimensional vector $\mathbf{z} \in \mathbb{R}^d$ by independently quantizing each dimension to a finite set of levels. For dimension i with L_i levels:

$$\text{FSQ}_i(z_i) = \text{round}\left(\frac{z_i \cdot (L_i - 1)}{2}\right) \cdot \frac{2}{L_i - 1} \tag{1}$$

The total codebook size is $\prod_{i=1}^{d} L_i$, with code rate $\sum_{i=1}^{d} \log_2(L_i)$ bits per token.

3.2. Naive Residual FSQ and Its Problems

A straightforward extension applies FSQ residually across K stages:

$$\mathbf{q}_1 = \text{FSQ}_1(\mathbf{z}), \quad \mathbf{r}_1 = \mathbf{z} - \mathbf{q}_1 \tag{2}$$

$$\mathbf{q}_k = \text{FSQ}_k(\mathbf{r}_{k-1}), \quad \mathbf{r}_k = \mathbf{r}_{k-1} - \mathbf{q}_k \tag{3}$$

A critical issue emerges in this naive approach: the residual magnitude decay problem. As quantization progresses, we observe that $\|\mathbf{r}_k\| \ll \|\mathbf{r}_{k-1}\|$, causing subsequent FSQ layers to operate on extremely weak signals. Empirically, we find residual magnitudes decay exponentially: $\|\mathbf{r}_k\| \approx \alpha^k \|\mathbf{z}\|$ where $\alpha < 0.3$. This severely limits their quantization effectiveness since FSQ's fixed boundaries become increasingly mismatched with the residual distribution. For instance, if FSQ boundaries are designed for unit-scale inputs, but $\mathbf{r}_3$ has magnitude 10^{-3}, the effective quantization resolution is reduced by three orders of magnitude.

3.3. Robust RFSQ Framework

To address residual decay, we propose two complementary conditioning strategies that maintain the simplicity of FSQ while enabling effective multi-stage quantization.

Our first strategy introduces learnable scaling factors that adaptively amplify residual signals before quantization:

$$\mathbf{q}_k = \text{FSQ}_k(\alpha_k \cdot \mathbf{r}_{k-1}) \tag{4}$$

$$\mathbf{r}_k = \mathbf{r}_{k-1} - \mathbf{q}_k / \alpha_k \tag{5}$$

where the scaling factors α_k are learned parameters initialized to 1.0. This approach allows each stage to adapt to the magnitude of its input while maintaining perfect reconstruction through inverse scaling.

The second strategy employs invertible layer normalization to stabilize the input distribution:

$$\hat{\mathbf{r}}_{k-1} = \text{LayerNorm}(\mathbf{r}_{k-1}) \tag{6}$$

$$\mathbf{q}_k = \text{FSQ}_k(\hat{\mathbf{r}}_{k-1}) \tag{7}$$

$$\mathbf{r}_k = \mathbf{r}_{k-1} - \text{LayerNorm}^{-1}(\mathbf{q}_k) \tag{8}$$

Algorithm 1 Robust Residual Finite Scalar Quantization

Require: Input $\mathbf{z} \in \mathbb{R}^{H \times W \times D}$, stages K, strategy
Ensure: Quantized features $\mathbf{q}_{\text{total}}$, indices $\{\mathbf{I}_1, \ldots, \mathbf{I}_K\}$

```
Initialize r_0 = z, q_total = 0
for k = 1 to K do
    if k = 1 then
        q_k, I_k = FSQ_k(r_{k-1}) {No conditioning}
        r_k = r_{k-1} − q_k
    else if strategy == "scale" then
        q_k, I_k = FSQ_k(α_k · r_{k-1})
        r_k = r_{k-1} − q_k/α_k
    else if strategy == "layernorm" then
        r̂_{k-1} = LayerNorm(r_{k-1})
        q_k, I_k = FSQ_k(r̂_{k-1})
        r_k = r_{k-1} − LayerNorm^{-1}(q_k)
    else
        q_k, I_k = FSQ_k(r_{k-1})
        r_k = r_{k-1} − q_k
    end if
    q_total = q_total + q_k
end for
return q_total, {I_1, ..., I_K}
```

This normalization ensures consistent input magnitudes across stages while preserving invertibility for perfect reconstruction.

Algorithm 1 presents the complete framework.

4. EXPERIMENTS

We evaluate RFSQ on two challenging modalities: large-scale audio reconstruction and image compression, demonstrating its effectiveness and generalizability. Code is available at `https://github.com/zhuxiaoxuhit/robust_rfsq`.

4.1. Audio Reconstruction Experiments

We evaluate RFSQ on a large-scale audio reconstruction task using the Emilia dataset [13], which provides diverse multilingual speech content. All audio is downsampled to 24kHz to balance computational efficiency with perceptual quality. Our encoder-decoder architecture follows the EnCodec [12] design principles, employing a compression ratio of 320 (from 24kHz to 75Hz latent representation) with 128-dimensional bottleneck features. This architecture has proven effective for high-quality neural audio compression.

All quantization methods are evaluated under a unified 24 bits/frame constraint to ensure fair comparison. The training objective combines multiple complementary loss terms that capture different aspects of audio quality:

Table 1: DNSMOS Evaluation Results (24 Bits/Frame)

Method	Config	DNSMOS	vs. RVQ
Traditional Baselines			
VQ-EMA-4×64-PQ	4×64 PQ	2.687 ± 0.468	-23.6%
LFQ-24D	24-dim	2.814 ± 0.437	-20.0%
FSQ-4D-Uniform	[64,64,64,64]	2.965 ± 0.383	-15.7%
State-of-the-Art			
RVQ-4×64	4-stage	3.518 ± 0.281	+0.0%
RFSQ Variants			
RFSQ-2S-NU-LN	2-stage LN	3.289 ± 0.296	-6.5%
RFSQ-4S-NU-No	4-stage None	3.187 ± 0.319	-9.4%
RFSQ-4S-NU-Scale	4-stage Scale	3.421 ± 0.274	-2.8%
RFSQ-8S-Uni-LN	8-stage LN	3.356 ± 0.262	-4.6%
RFSQ-4S-Uni-LN	4-stage Uni	3.598 ± 0.258	+2.3%
RFSQ-4S-NU-LN	**4-stage LN**	$\mathbf{3.646 \pm 0.251}$	$+\mathbf{3.6\%}$

$$\mathcal{L} = \lambda_{\text{time}} \|\mathbf{x} - \hat{\mathbf{x}}\|_1 + \lambda_{\text{stft}} \mathcal{L}_{\text{STFT}} + \lambda_{\text{spec}} \mathcal{L}_{\text{spec}} + \lambda_{\text{adv}} \mathcal{L}_{\text{adv}} + \lambda_{\text{feat}} \mathcal{L}_{\text{feat}} \tag{9}$$

where $\lambda_{\text{time}} = 1.0$, $\lambda_{\text{stft}} = 1.0$, $\lambda_{\text{spec}} = 0.1$, $\lambda_{\text{adv}} = 1.0$, and $\lambda_{\text{feat}} = 2.0$. These weights were determined through extensive validation to balance reconstruction fidelity with perceptual quality.

We compare RFSQ against a comprehensive set of baselines and variants. The traditional baselines include vector quantization with exponential moving average using product quantization (VQ-EMA-4×64-PQ: 4 subvectors with codebook size 64 each), lookup-free quantization (LFQ-24D: 24-dimensional learnable quantization), and single-stage FSQ (FSQ-4D-Uniform: 4 dimensions with levels=[64,64,64,64]). The state-of-the-art baseline is residual vector quantization (RVQ-4×64: 4 stages with codebook size 64 each).

For RFSQ, we evaluate six carefully designed variants. The 4-stage non-uniform configuration (RFSQ-4S-NU-LN) uses a "front-heavy" bit allocation: Stage 1 with 8 bits (levels=[16,16]), Stage 2 with 6 bits (levels=[8,8]), and Stages 3-4 with 5 bits each (levels=[8,4]). This design allocates more bits to early stages where signal magnitude is highest. We compare this against uniform allocation (RFSQ-4S-Uni-LN: 4 stages with 6 bits each, levels=[8,8]) to validate our allocation strategy. The conditioning strategy ablation includes LayerNorm (RFSQ-4S-NU-LN), Scale (RFSQ-4S-NU-Scale), and None (RFSQ-4S-NU-No) variants. Finally, we explore different granularities: fine-grained 8-stage (RFSQ-8S-Uni-LN: 8 stages with 3 bits each, levels=[4,2]) and coarse-grained 2-stage (RFSQ-2S-NU-LN: Stage 1 with 14 bits using 3D levels=[32,32,16], Stage 2 with 10 bits using levels=[16,16,4]).

Table 1 and Figure 1 present the DNSMOS [14] evaluation results across all methods. The results reveal several important insights about the effectiveness of our proposed approach.

RFSQ-4S-NU-LN achieves DNSMOS 3.646, a 3.6% improvement over RVQ (3.518), with only 4.3% degradation from original audio (3.810). The non-uniform bit allocation (8, 6, 5, 5 bits per stage) effectively captures signal energy distribution, where early stages encode dominant components while later stages refine details. LayerNorm conditioning provides a substantial 14.4% gain over unconditioned variants (3.646 vs. 3.187), empirically validating our theoretical analysis of residual magnitude decay. Scale conditioning achieves intermediate performance (3.421), suggesting that adaptive normalization better addresses cross-stage magnitude variations than simple scaling.

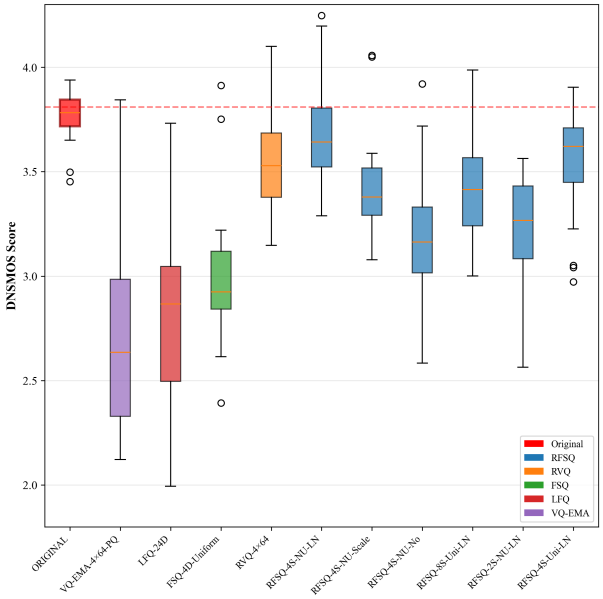

Fig. 1: DNSMOS evaluation results. Box plots show score distributions, with RFSQ variants (blue) consistently outperforming traditional baselines. Red dashed line indicates original audio quality (3.810).

The 4-stage configuration emerges as optimal, balancing quantization precision with error accumulation. Fewer stages (2-stage: 3.289) lack sufficient representational capacity, while excessive stages (8-stage: 3.356) accumulate quantization errors that degrade overall quality. This finding aligns with our theoretical framework: each stage introduces bounded error ϵ_k, and total error grows as $O(\sqrt{K})$ for K stages.

All single-stage baselines exhibit severe performance degradation: VQ-EMA (-23.6%), LFQ (-20.0%), FSQ (-15.7%). This consistent gap demonstrates the fundamental limitation of single-stage quantization in capturing the multi-scale nature of audio signals, validating the necessity of residual decomposition for high-quality reconstruction under tight bit constraints.

Table 2: RFSQ Performance on ImageNet

Method	Bits	L1↓	LPIPS↓	PSNR↑
22.0-bit Configurations				
RFSQ-2×2048-None	22.0	0.130	0.159	21.1
RFSQ-2×2048-Scale	22.0	0.122	0.152	21.5
RFSQ-2×2048-LN	22.0	0.124	0.148	21.3
40.0-bit Configurations				
RFSQ-4×1024-None	40.0	0.113	0.121	22.2
RFSQ-4×1024-Scale	40.0	0.103	0.101	22.9
RFSQ-4×1024-LN	**40.0**	**0.102**	**0.100**	**22.9**

4.2. Image Reconstruction Experiments

To demonstrate the generalizability of RFSQ beyond audio applications, we conduct comprehensive experiments on image reconstruction using the ImageNet dataset [15]. We evaluate six RFSQ configurations at 128×128 resolution, systematically comparing different bit rates (22.0 and 40.0 bits) and conditioning strategies to understand their impact on visual quality.

Our encoder-decoder architecture employs a symmetric design with two downsampling layers in the encoder using 4×4 convolutional kernels with stride 2, followed by a 1×1 convolution for dimension adjustment, transforming the 128×128 input to 32×32 feature representations. The decoder mirrors this structure with transposed convolutions for upsampling. We employ a combination of L1 reconstruction loss and LPIPS perceptual loss [16]:

$$\mathcal{L} = \lambda_1 \|\mathbf{x} - \hat{\mathbf{x}}\|_1 + \lambda_p \text{LPIPS}(\mathbf{x}, \hat{\mathbf{x}}) \tag{10}$$

where $\lambda_1 = \lambda_p = 1.0$ for equal weighting between pixel-level accuracy and perceptual quality.

Figure 2 shows 40-bit RFSQ achieves superior detail preservation over 22-bit variants. The bit budget increase (22→40) provides larger gains than conditioning strategies alone, though conditioning remains crucial at lower bit rates.

Table 2 reveals a nuanced relationship between bit budget and conditioning strategies. While increasing bits from 22 to 40 yields substantial improvements (e.g., L1 loss reduction from 0.113 to 0.102 for LayerNorm), the choice of conditioning strategy can achieve comparable gains without additional bits. At 40 bits, LayerNorm conditioning improves L1 loss by 9.7% (0.102 vs. 0.113) and perceptual loss by 17.4% (0.100 vs. 0.121) compared to no conditioning, demonstrating that intelligent residual processing is as critical as raw capacity.

5. CONCLUSION

We propose Robust Residual Finite Scalar Quantization (RFSQ), a novel framework addressing the fundamental residual magnitude decay problem through intelligent conditioning strategies. Our experiments demonstrate RFSQ's effectiveness across modalities: 3.6% DNSMOS improvement in audio and significant gains in image quality (9.7% L1, 17.4% perceptual improvement at 40 bits).

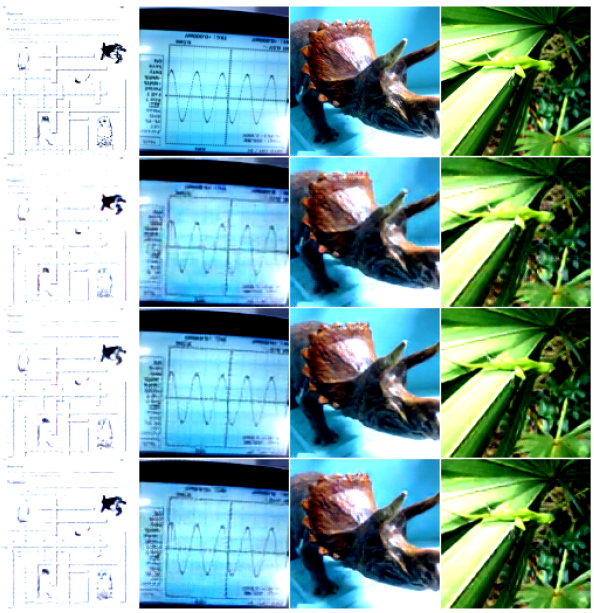

Fig. 2: Visual quality comparison. From top: original, RFSQ-2×2048-LN (22.0 bits), RFSQ-4×1024-LN (40.0 bits), RFSQ-4×1024-None (40.0 bits).

The LayerNorm strategy's superiority stems from maintaining consistent input statistics across stages, counteracting exponential magnitude decay. By normalizing residuals before each quantization stage, RFSQ ensures all stages contribute meaningfully, unlike traditional methods where later stages encode mainly noise. This addresses the core challenges of magnitude decay and inter-stage dependencies. The consistent improvements across diverse datasets and modalities validate RFSQ's robustness and broad applicability. RFSQ serves as a plug-and-play module for neural compression, combining FSQ's simplicity with multi-stage quantization's effectiveness. Future work includes adaptive stage allocation and video compression applications.

6. ACKNOWLEDGEMENTS

The audio encoder/decoder architectures are based on the SEANet design from EnCodec [12]. The image reconstruction experiments were inspired by duchenzhuang's FSQ-pytorch project. We also acknowledge the EnCodec codebase for providing reference implementations that guided our audio reconstruction experiments.

7. REFERENCES

[1] Robert M Gray, "Vector quantization," *IEEE ASSP Magazine*, vol. 1, no. 2, pp. 4–29, 1984.

[2] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu, "Neural discrete representation learning," in *31st Conference on Neural Information Processing Systems (NIPS 2017)*, Long Beach, CA, USA, 2017.

[3] Ali Razavi, Aaron van den Oord, and Oriol Vinyals, "Generating diverse high-fidelity images with vq-vae-2," in *33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, Vancouver, Canada, 2019.

[4] Fabian Mentzer, David Minnen, Eirikur Agustsson, and Michael Tschannen, "Finite scalar quantization: Vq-vae made simple," *arXiv preprint arXiv:2309.15505*, 2023.

[5] Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, Fan Yu, Huadai Liu, Zhengyan Sheng, Yue Gu, Chong Deng, Wen Wang, Shiliang Zhang, Zhijie Yan, and Jingren Zhou, "Cosyvoice 2: Scalable streaming speech synthesis with large language models," *arXiv preprint arXiv:2412.10117*, 2024.

[6] Wei Deng, Siyi Zhou, Jingchen Shu, Jinchao Wang, and Lu Wang, "Indextts: An industrial-level controllable and efficient zero-shot text-to-speech system," *arXiv preprint arXiv:2502.05512*, 2025.

[7] Kunal Dhawan, Nithin Rao Koluguri, Ante Jukić, Ryan Langman, Jagadeesh Balam, and Boris Ginsburg, "Codec-asr: Training automatic speech recognition on discrete speech representations," in *Proc. Interspeech*, 2024.

[8] Julian D Parker, Anton Smirnov, Jordi Pons, CJ Carr, Zack Zukowski, Zach Evans, and Xubo Liu, "Scaling transformers for low-bitrate high-quality speech coding," *arXiv preprint arXiv:2411.19842*, 2024.

[9] Adrian Łańcucki, Jan Chorowski, Guillaume Sanchez, Ricard Marxer, Nanxin Chen, Hans J.G.A. Dolfing, Sameer Khurana, Tanel Alumäe, and Antoine Laurent, "Robust training of vector quantized bottleneck models," in *International Joint Conference on Neural Networks (IJCNN)*, 2020.

[10] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman, "Maskgit: Masked generative image transformer," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

[11] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi, "Soundstream: An end-to-end neural audio codec," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

[12] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi, "High fidelity neural audio compression," *arXiv preprint arXiv:2210.13438*, 2022.

[13] Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, Yuancheng Wang, Kai Chen, Pengyuan Zhang, and Zhizheng Wu, "Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation," in *IEEE Spoken Language Technology Workshop (SLT)*, 2024.

[14] Babak Naderi and Ross Cutler, "An open source implementation of itu-t recommendation p.808 with validation," in *Proc. Interspeech*, 2020, pp. 2862–2866.

[15] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[16] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang, "The unreasonable effectiveness of deep features as a perceptual metric," *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.