VRVQ: Variable Bitrate Residual Vector Quantization for Audio Compression

Sony AI
SONY

Yunkee Chae[1,3*], Woosung Choi[1], Yuhta Takida[1], Junghyun Koo[1], Yukara Ikemiya[1], Zhi Zhong[2], Kin Wai Cheuk[1], Marco A. Martínez-Ramírez[1], Kyogu Lee[3,4,5], Wei-Hsiang Liao[1], Yuki Mitsufuji[1,2]

[1]Sony AI, [2]Sony Group Corporation, Tokyo, Japan, [3]IPAI, [4]AIIS, [5]Department of Intelligence and Information, Seoul National University

MARG Music & Audio Research Group

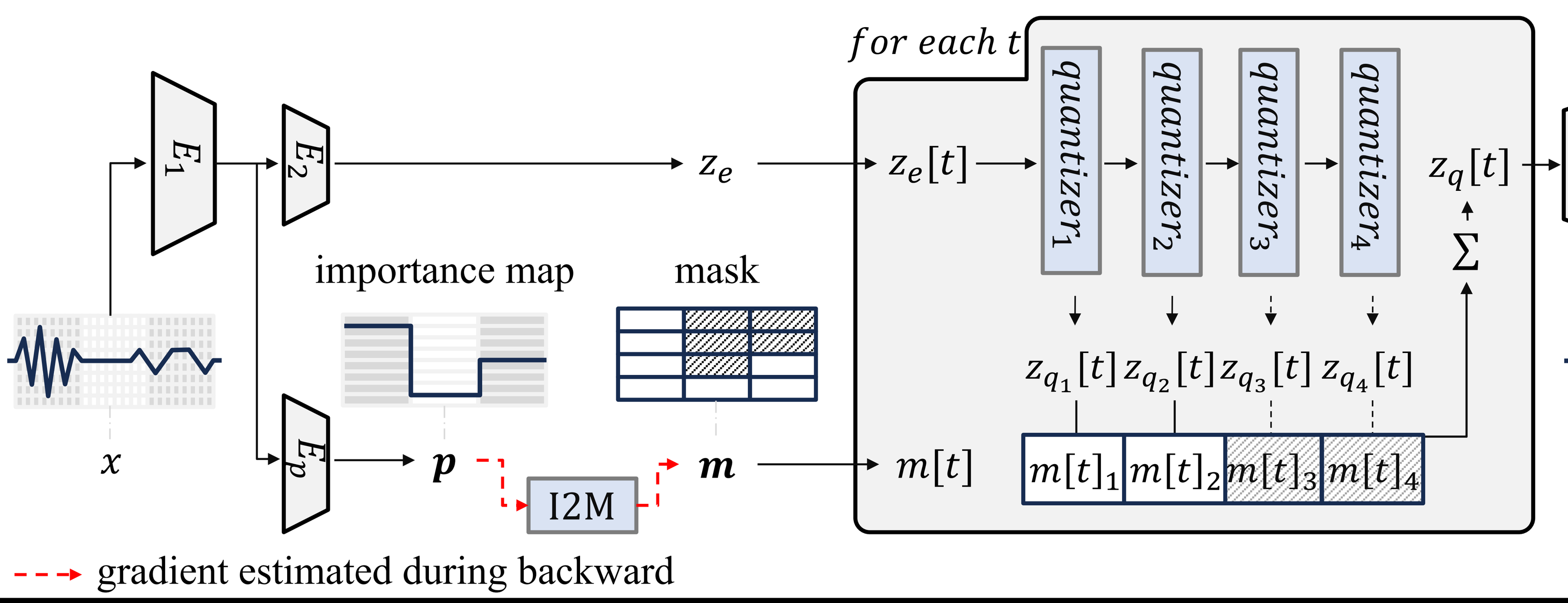
Motivation

- Recent SOTA neural audio codecs have adopted residual vector quantization (RVQ).
- The current RVQ codec uses the same number of codebooks for each time frame.
- In other words, once the target bandwidth is set, it allocates a **constant bitrate (CBR)** across all frames
- CBR can lead to a waste of bitrate in frames with low information content, such as silence.

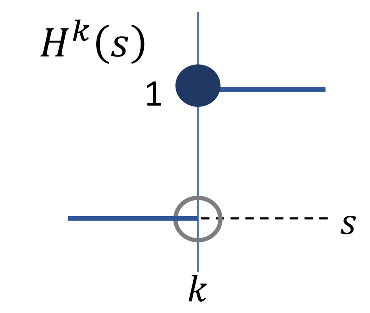
Contribution

- We propose Variable Bitrate (VBR) RVQ framework
  : An RVQ that allocates different bitrates to each frame by using different number of codebooks per frame.
- We apply VBR scheme to RVQ (or RVQGAN) for the first time.
- We base our approach on ***importance map***, which has been employed in image compression.
- We identify issues with existing training methods and propose an improved approach to enhance the gradient flow.

VBR RVQ with Importance Map



- We train audio codec jointly with importance subnet
- $p \in (0,1)^T$
- Optimize rate-distortion tradeoff: $\boldsymbol{\mathcal{L} = \mathcal{L}_D + \beta \mathcal{L}_R}$
- $\mathcal{L}_D$: reconstruction loss used in the existing RVQGAN.
- $\mathcal{L}_R = \frac{1}{T}\sum_{t=1}^{T}|p[t] - 0| = \frac{1}{T}\sum_{t=1}^{T} p[t]$
- $I2M(p[t]) = \left[H^0(S(p[t])), \dots, H^{N_q-1}(S(p[t]))\right]$
  - $S(p) = N_q \cdot p \in (0, N_q)$
- **This operation is non-differentiable**



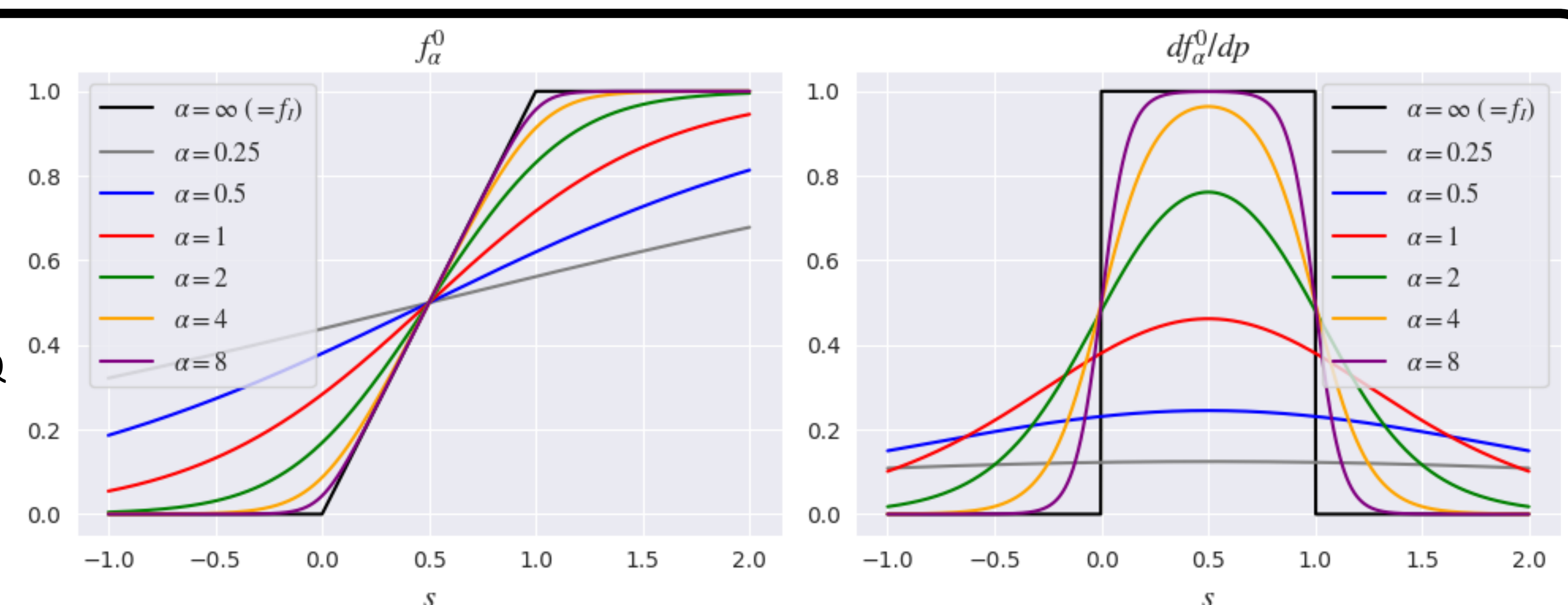
Smoothing the Surrogate Function

- We define the "surrogate" of the H^k for the backpropagation:
- We use the straight-through estimation (STE)

$$x \mapsto I2M_{soft}(x) + sg\left(I2M(x) - I2M_{soft}(x)\right)$$

- Previous work [1] in the image compression model used:
  $f_I^k = \max(\min(s-k, 1), 0)$ (i.e., "identity for the backward pass")
- f_I makes the model suboptimal and degrades the performance of VRVQ
  - Gradient does not flow through large regions due to *max* and *min* ops.
  - Non-zero gradient can exist for only a single $k \in \{0, \cdots, N_q - 1\}$
- To address this, we propose a smooth surrogate function

$$f_\alpha^k(s) = \frac{1}{2\alpha}\log\left(\frac{\cosh(\alpha(s-k))}{\cosh(\alpha(-s+k+1))}\right) + \frac{1}{2}$$



$$\Rightarrow I2M_{soft}(p[t]) = \left[f_\alpha^0(S(p[t])), \dots, f_\alpha^{N_q-1}(S(p[t]))\right]$$

Random Scaling for Rate Control

- **Previous works:**
  - Once model is trained, **importance map p is fixed** for the input x.
  - This restricts the flexibility of the model in terms of **rate control** within a single model.
- **Proposed:**
  - Meanwhile, RVQ-based models control the rate using structured dropout
    - with **random sampling** of number of codebook $n_q \in \{1, \dots, N_q\}$.
  - We base our approach on the importance map and incorporate random scaling
    - allowing a single model to support **multiple bitrates.**
- **Random Scaling:** $S(p) = l \cdot p$, where $l \sim Uni([L_{min}, L_{max}])$

Dataset / Setups

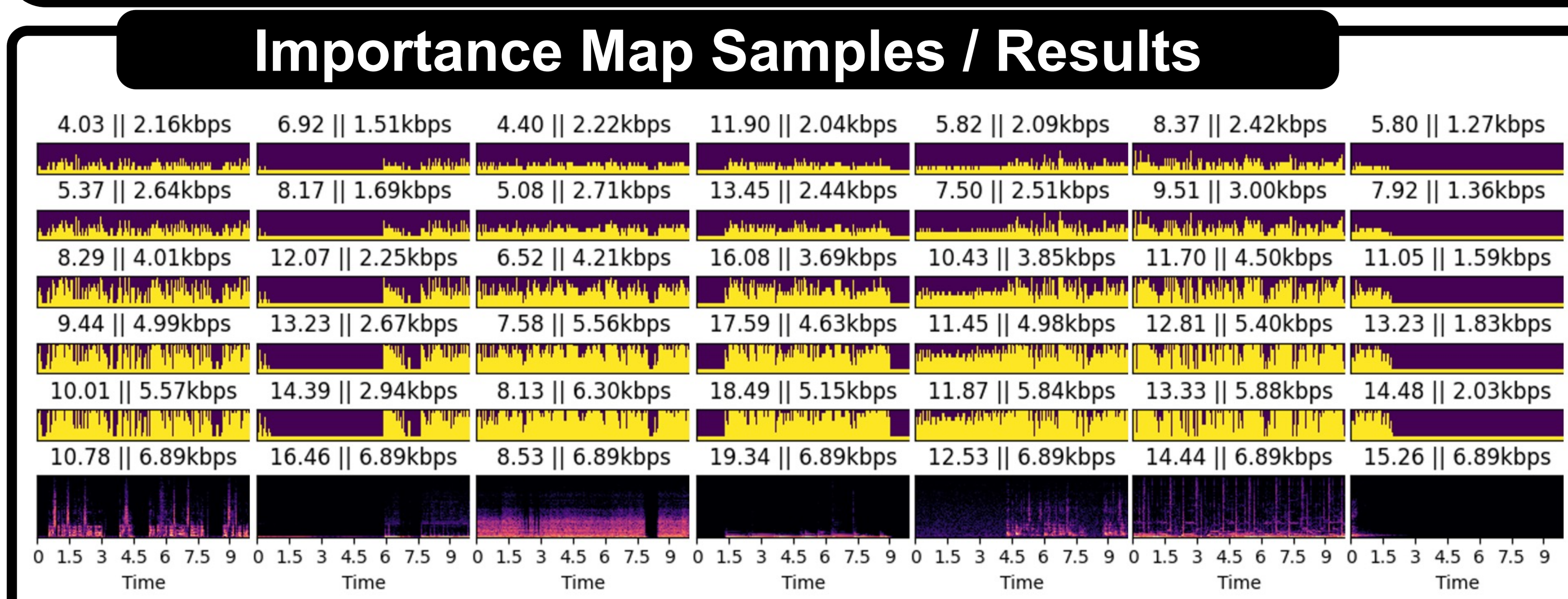
Dataset

- **Train Set**
  - Speech: DAPS, CommonVoice, VCTK
  - Music: MUSDB18, MTG-Jamendo
  - General: AudioSet
- **Eval Set**
  - Speech: DAPS test (F10, M10)
  - Music: MUSDB18 test
  - General: AudioSet eval

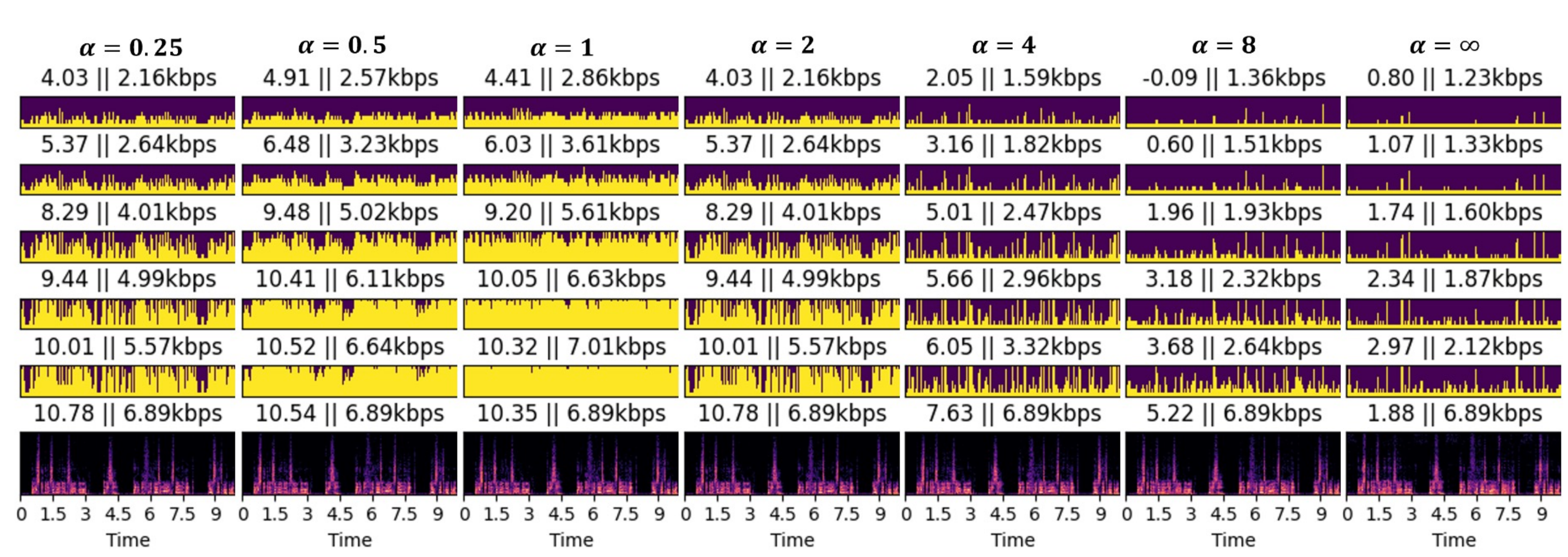
Setups

- Codec: DAC [2] with $N_q = 8$
  - Due to transmission cost $\lceil \log_2 N_q \rceil$ of VRVQ
  - i.e., +0.238 kbps for bitrate calculation.
- Rate loss weight $\beta = 2$
- Batch size: 32
- Train iteration: 300k for each exp.
- $L_{min} = 1,\ L_{max} = 48$

Importance Map Samples / Results



(a) Codebook usage based on importance map with varying l for seven audio samples when $\alpha = 2$. The bottom row shows the spectrogram of the input audio.



(b) Codebook usage based on importance map for the same sample with varying α values. The bottom row shows the spectrogram of the reconstructed audio with the full number of codebooks.

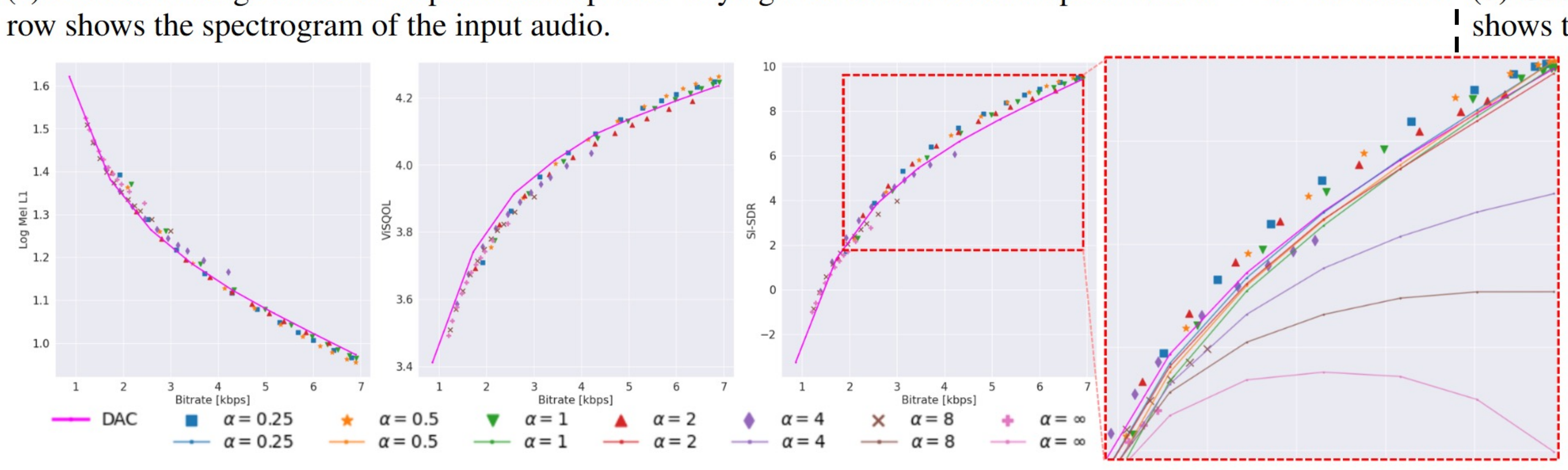


Figure 2: The results of VRVQ across different α. The points marked with various markers represent the results of inference at different scaling factor, l=4, 6, 8, 10, 12, 14, 16, 18, 20, 24, and 32, in VBR mode. In the rightmost plot, we display solid lines representing the results of inference in CBR mode for each model.

Importance Map

- Each row of the importance maps denotes the **level l** from 4 to 26.
- For silence, importance map decides to use only one codebook, regardless how high the importance map is scaled.
- As alpha increases, the importance map becomes more spiky and uses fewer codebooks, and at the base lien f_I, it doesn't utilize many codebooks.

Rate-Distortion (RD) Curves

- **Solid line** refers to the results of **CBR mode** of our models: simply ignoring importance map and using a constant number of codebooks for all frames.
- The performance (RD-curve) degrades as α increases.
- When $\alpha \leq 2$, performs better in RD compared to DAC.

References

[1] F. Mentzer, E. Agustsson, M. Tschannen, R. Timofte, and L. Van Gool, "Conditional probability models for deep image compression," in *CVPR 2018*
[2] R. Kumar, P. Seetharaman, A. Luebs, I. Kumar, and K. Kumar, "High-fidelity audio compression with improved rvqgan," in NeurIPS 2023