Fraunhofer HHI

mhv/2024

UNIVERSITÄT KLAGENFURT

Video Quality Assessment with Texture Information Fusion for Streaming Applications

Vignesh V Menon[1], Prajit T Rajendran[2], Reza Farahani[3], Klaus Schoeffmann[3], Christian Timmerer[1]

[1]Video Communication and Applications Dept., Fraunhofer HHI, Berlin, Germany
[2]CEA, List, F-91120 Palaiseau, Université Paris-Saclay, France
[3]Christian Doppler Laboratory ATHENA, Alpen-Adria-Universität, Klagenfurt, Austria

12 Feb 2024

Outline

Introduction

Reduced-reference VQA (RR-VQA)

- With all the available encoding options and trade-offs to consider in *HTTP Adaptive Streaming* (HAS),[1] having a lightweight and reliable VQA method is crucial.
- The advantage of `RR-VQA` lies in its ability to evaluate video quality using limited information, making it more suitable for real-time VQA, especially in adaptive streaming or live broadcast scenarios.[2]

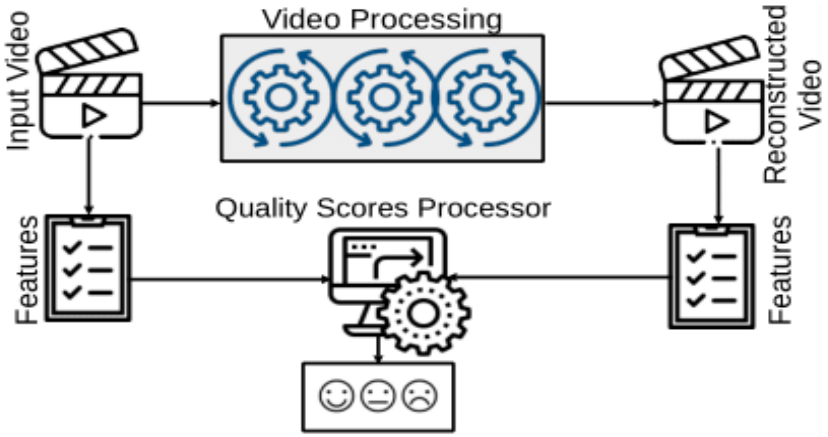


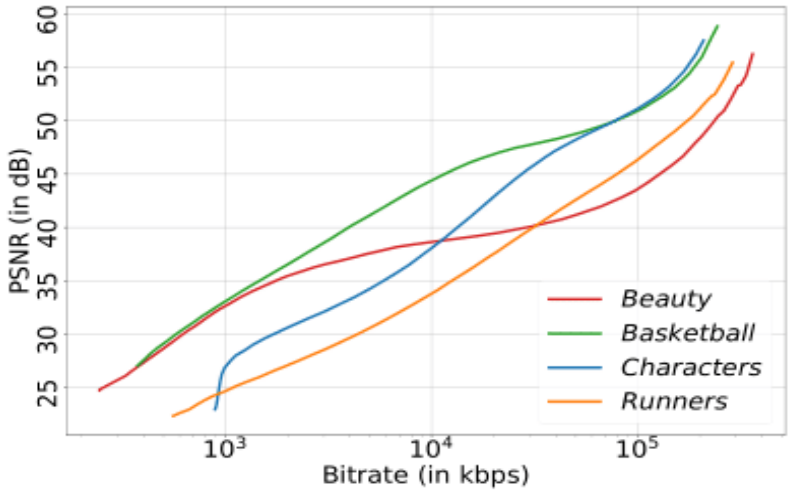
Figure: The structure of state-of-the-art `RR-VQA` methods utilized, especially within streaming video coding systems.

[1] Abdelhak Bentaleb et al. "A Survey on Bitrate Adaptation Schemes for Streaming Media Over HTTP". In: *IEEE Communications Surveys Tutorials* 21.1 (2019), pp. 562–585. DOI: 10.1109/COMST.2018.2862938. URL: https://doi.org/10.1109/COMST.2018.2862938.

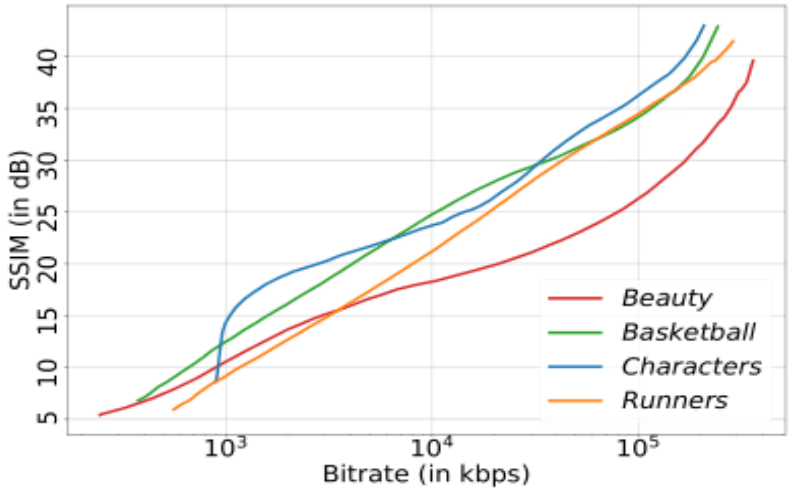
[2] Vignesh V Menon et al. *Content-Adaptive Variable Framerate Encoding Scheme for Green Live Streaming*. 2023. arXiv: 2311.08074 [cs.MM].

Introduction

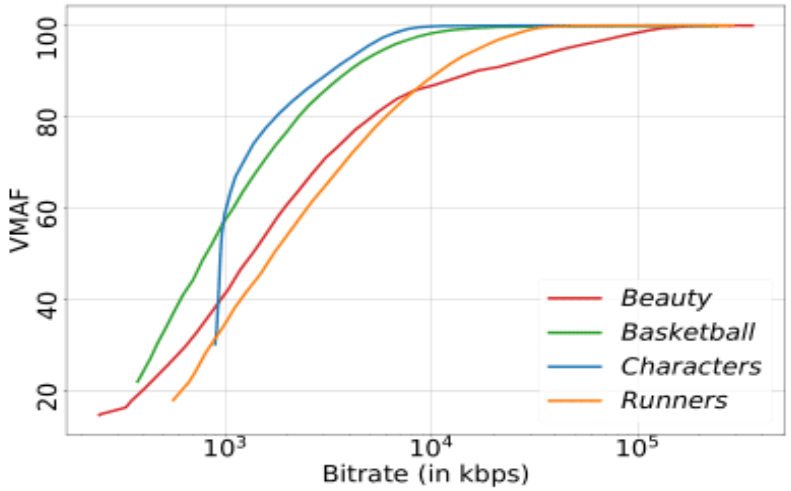
Correlation between VQA metrics



(a) PSNR



(b) SSIM



(c) VMAF

Figure: Rate-distortion (RD) curves of selected segments of different spatiotemporal complexities – *Beauty* (E_Y =59.90, h =17.49, L_Y =89.25), *Basketball* (E_Y =15.30, h =12.59, L_Y =108.18), *Characters* (E_Y =45.42, h =36.88, L_Y =134.56), and *Runners* (E_Y =105.85, h =22.48, L_Y =126.60). The segments are downsampled to 30 fps and encoded with the x264 AVC encoder using *ultrafast* preset and CRF rate control.

Introduction

Table: Pearson Correlation of VQA metrics.

Metric	PSNR	SSIM	VMAF
PSNR	1.00	0.70	0.83
SSIM	0.70	1.00	0.88
VMAF	0.83	0.88	1.00

Target

- The expected computation time should be comparable to PSNR and SSIM, with the highest possible accuracy compared to the VMAF score.
- The proposed VQA method is expected to replace the state-of-the-art VMAF computation in streaming applications.

VQ-TIF

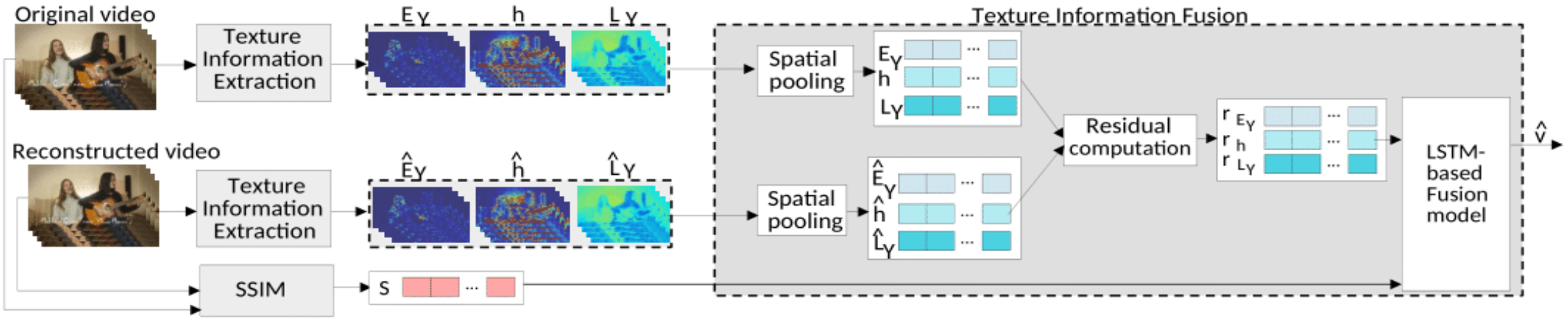


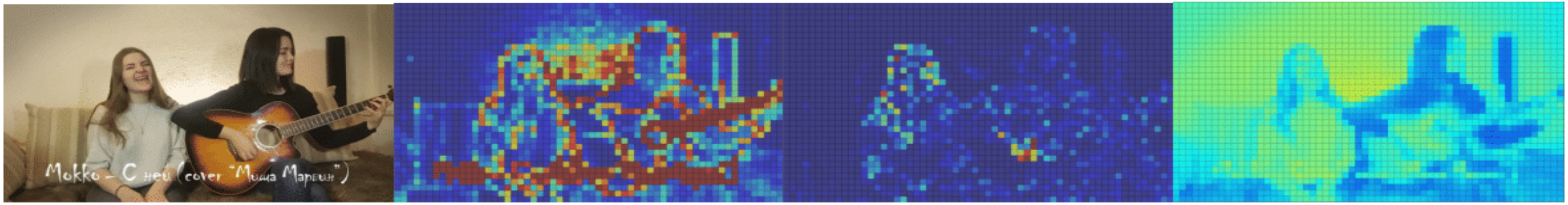
Figure: VQA for a video segment using VQ-TIF model envisioned in this paper.

1. frame-wise *texture information extraction* for each chunk
2. SSIM calculation
3. *texture information fusion*, where the features and the computed SSIM are fused using an LSTM-based model to determine the VQ-TIF score for each chunk

VQ-TIF

Texture information extraction

Three DCT-energy-based features, the average luma texture energy E_Y, the average gradient of the luma texture energy h, and the average luminescence L_Y are used as the texture information measures.[3,4]



(a) original frame (b) E_Y (c) h (d) L_Y

Figure: Heatmap depiction of the luma texture information $\{E_Y, h, L_Y\}$ extracted from the second frame of *CoverSong_1080P_0a86* video of Youtube UGC Dataset.[5]

[3] Vignesh V Menon et al. "Green Video Complexity Analysis for Efficient Encoding in Adaptive Video Streaming". In: *Proceedings of the First International Workshop on Green Multimedia Systems*. 2023, 16–18. ISBN: 9798400701962. DOI: 10.1145/3593908.3593942. URL: https://doi.org/10.1145/3593908.3593942.

[4] Vignesh V Menon et al. "JND-aware Two-pass Per-title Encoding Scheme for Adaptive Live Streaming". In: *IEEE Transactions on Circuits and Systems for Video Technology* (2023), pp. 1–1. DOI: 10.1109/TCSVT.2023.3290725. URL: https://doi.org/10.1109/TCSVT.2023.3290725.

[5] Yilin Wang, Sasi Inguva, and Balu Adsumilli. "YouTube UGC Dataset for Video Compression Research". In: *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*. Sept. 2019. DOI: 10.1109/mmsp.2019.8901772. URL: https://doi.org/10.1109/mmsp.2019.8901772.

VQ-TIF

Spatial pooling

The video segments are divided into T chunks with a fixed number of frames (*i.e.*, f_c). The averages of the E_Y, h and L_Y features of each frame in the chunk are calculated to obtain the spatially pooled representation of the chunk, expressed as:

$$X = \{x_1, x_2, .., x_{f_c}\}, \tag{1}$$

$$\hat{X} = \{\hat{x}_1, \hat{x}_2, .., \hat{x}_{f_c}\}, \tag{2}$$

where x_i and $\hat{x}_i$ are the i^{th} frame feature set associated with the original and reconstructed video chunks, respectively.

$$x_i = [E_i, h_i, L_i], \tag{3}$$

$$\hat{x}_i = [\hat{E}_i, \hat{h}_i, \hat{L}_i] \quad \forall i \in [1, f_c] \tag{4}$$

VQ-TIF

Residual computation

Residual features are formed by subtracting the original video texture information features from the reconstructed video features. This difference is known as the error or residual feature, expressed as:

$$r_{E_i} = E_i - \hat{E}_i \tag{5}$$

$$r_{h_i} = h_i - \hat{h}_i \tag{6}$$

$$r_{L_i} = L_i - \hat{L}_i \tag{7}$$

where $i \in [1, f_c]$.

The residual features usually have low information entropy, as the original and reconstructed video frames are similar. The entropy increases with increased distortion introduced in the reconstructed video.

VQ-TIF

Fusion

- The fusion of the texture information features is established using a *long short-term memory* (LSTM).
- Frame-wise SSIM values denoted by $S = \{s_1, s_2, .., s_{f_c}\}$ are appended to the residual features.
- The prediction model is a function of the residual features of the frames and the SSIM values in a chunk, as shown below:

$$\tilde{x}_i = [r_i|s_i]^T \quad i \in [1, f_c] \tag{8}$$

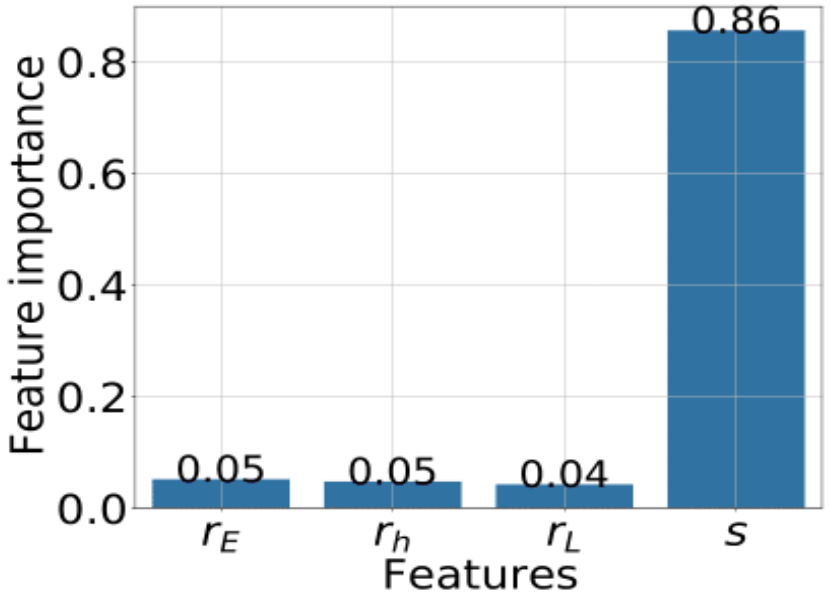
where $r_i = [r_{E_i}, r_{h_i}, r_{L_i}]$.

The estimated VQ-TIF score per chunk $\hat{v}$ can be presented as:

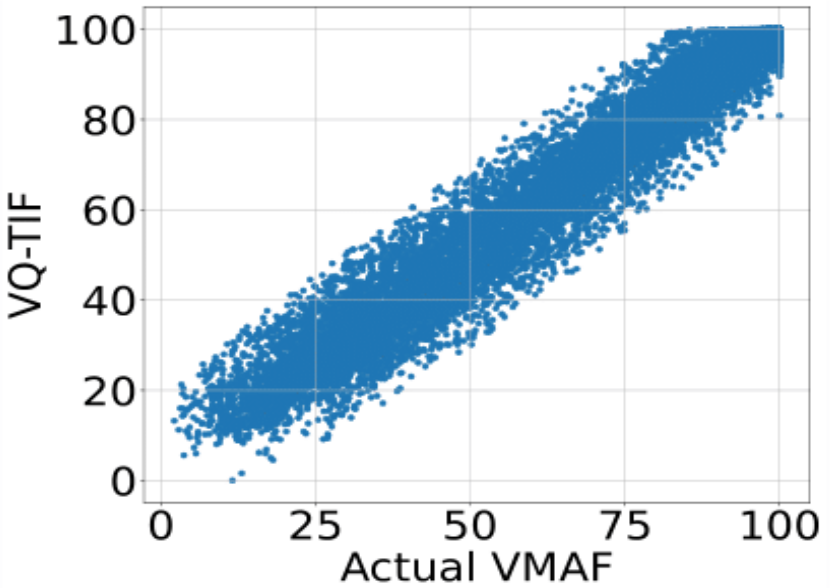
$$\hat{v} = f(\tilde{x}). \tag{9}$$

The VQ-TIF score of the reconstructed video segment is the average of the $\hat{v}$ values estimated for every chunk.

Results



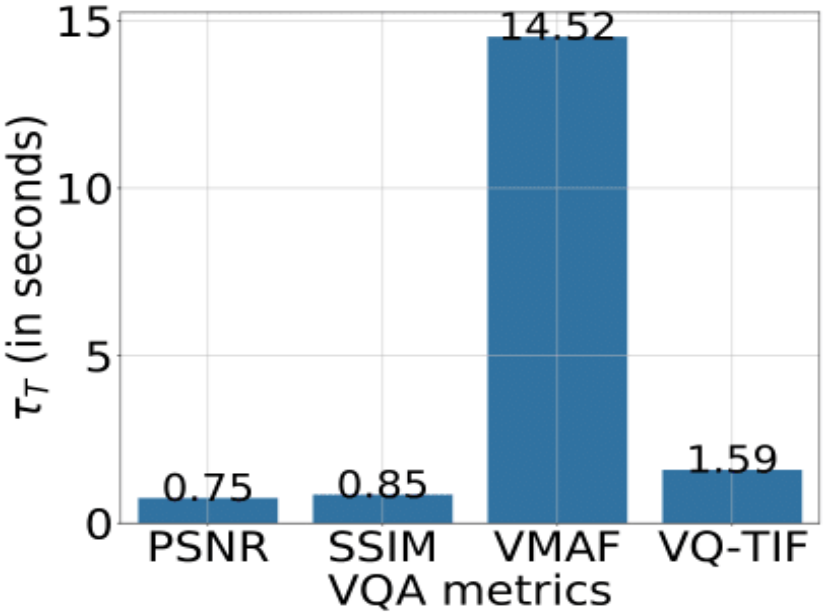
(a) Univariate feature importance of the LSTM model



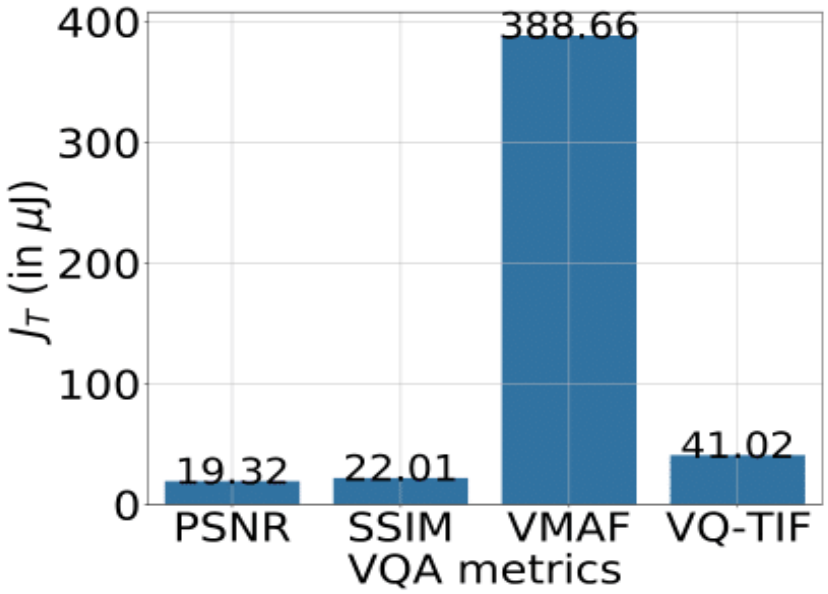
(b) Scatterplot of VMAF and `VQ-TIF`

- SSIM feature contributes the most to the `VQ-TIF` estimation, followed by r_E, r_h, and r_L features.
- The average PCC of the `VQ-TIF` scores to the VMAF score in the evaluation dataset is 0.96, while MAE is 2.71.

Results



(a) Total processing time (τ_T)



(b) Total energy consumption (J_T)

- The computation speed of `VQ-TIF` is 9.14 times higher than the state-of-the-art VMAF evaluation.
- In terms of total energy consumption, `VQ-TIF` saves 89.44 % compared to the state-of-the-art VMAF implementation.

Conclusions

- We proposed VQ-TIF, a fast and accurate reduced-reference video quality assessment (RR-VQA) method based on texture information fusion.
- VQ-TIF includes DCT-energy-based video complexity feature extraction where features representing luma texture and temporal activity are extracted from the original and reconstructed video segments.
- The extracted texture information is fused using an LSTM-based model to determine the VQ-TIF score.
- VQ-TIF is determined at a speed of 9.14 times faster than the state-of-the-art implementation of VMAF for Ultra HD (2160p) videos, consuming 89.44 % less energy. At the same time, VQ-TIF scores yield a PCC of 0.96 and MAE of 2.71 compared to the VMAF scores.

Limitations and future directions

- The evaluation of the proposed `VQ-TIF` model is limited to static dynamic range (SDR) content.
- `VQ-TIF` model can be extended to determine visual quality at multiple resolutions, including 8K (4320p).
- Various signal distortions may be considered during the model training to enhance the application scope.

Q & A

Thank you for your attention!

Vignesh V Menon (vignesh.menon@hhi.fraunhofer.de)