HoloQA: Full Reference Video Quality Assessor of Rendered Human Avatars in Virtual Reality

Avinab Saha, Yu-Chih Chen, Christian Häne, Jean-Charles Bazin, Ioannis Katsavounidis, *Senior Member, IEEE*, Alexandre Chapiro, and Alan C. Bovik, *Fellow, IEEE*

***Abstract*—We present HoloQA, a new state-of-the-art Full Reference Video Quality Assessment (VQA) model that was designed using principles of visual neuroscience, information theory, and self-supervised deep learning to accurately predict the quality of rendered digital human avatars in Virtual Reality (VR) and Augmented Reality (AR) systems. The growing adoption of VR/AR applications that aim to transmit digital human avatars over bandwidth-limited video networks has driven the need for VQA algorithms that better account for the kinds of distortions that reduce the quality of rendered and viewed avatars. As we will show, standard VQA models often fail to capture distortions unique to the rendering, transmission, and compression of videos containing human avatars. Towards solving this difficult problem, we adopt a multi-level *Mixture-of-Experts* approach. This involves computing distortion-aware perceptual features and high-level content-aware deep features that capture semantic attributes of human body avatars. The high-level features are computed using a self-supervised, pre-trained deep learning network. We show that HoloQA is able to achieve state-of-the-art performance on the recently introduced LIVE-Meta Rendered Human Avatar VQA database, demonstrating its efficacy in predicting the quality of rendered human avatars in VR. Furthermore, we demonstrate the competitive performance of HoloQA on other digital human avatar databases and on another synthetically generated video quality use case: cloud gaming. The code associated with this work will be made available on GitHub.**

***Index Terms*—Virtual Reality, Augmented Reality, Rendered Human Avatars, Full Reference VQA, Video Quality Assessment**

I. Introduction

In recent years, there have been significant strides in head-mounted displays (HMDs) and virtual/augmented reality (VR/AR) technologies. These advances have caught the attention of millions of consumers, researchers in academia, and large technology companies like Meta Platforms, Google, Apple, and Microsoft. Substantial investments are being made in this burgeoning industry, driven by expectations for future increases in the adoption of VR and AR. According to a report by Grand View Research [1], the global VR/AR market size was estimated at $59.96 Billion in 2022 and is expected to grow at a compound annualized growth rate of 27.5% from 2023 to 2030.

VR/AR technologies have already evolved rapidly with a focus on creating immersive, realistic, and comfortable experiences. Today, VR/AR technologies are used in domains such as virtual telepresence, online collaboration, cloud gaming and streaming media, healthcare, remote education, and tourism [2], [3]. These new technologies provide users with an immersive medium that allows them to observe and actively engage with real or rendered content. One central application of VR/AR is telepresence - that is, the presentation of a digital human element. These reproductions may be lifelike 3D models or stylized digital replicas (often called "avatars" or "Holograms," albeit technically incorrect, given that holography is not involved.). Having the capacity to communicate with others in an immersive virtual environment is an appealing technological prospect, which, if successful, will enable users to engage in visually intimate personal connections remotely. Immersive digital spaces where users are able to communicate and collaborate in ways that mimic real-world interactions while benefiting from the convenience and possibilities of the virtual world are expected to become more available due to significant recent improvements in deep learning and graphics for digital human rendering.

Motivated by these developments, this work targets telepresence scenarios where users engage in remote visual interactions within virtual, yet realistic environments. Since human avatars are our focus, we are especially interested in the faithful emulation of facial expressions, hand and body movements, as their accurate reproduction is needed to ensure realism. Early-stage attempts to create such communication platforms include products like Meta's Horizon Meeting Rooms [4]. As these technologies mature, the progression of human avatar or "digital human" representations toward heightened degrees of realism and immersion, has been accompanied by notable increases in data volume. This issue may increase due to increases in spatial and temporal display resolution. Due to limits on available networks, computing, and device bandwidths, volumetric and texture compression

This work was supported by Meta Platforms, Inc. A.C. Bovik was supported in part by the National Science Foundation AI Institute for Foundations of Machine Learning (IFML) under Grant 2019844.

Avinab Saha, and Alan C Bovik are with the Department of Electrical and Computer Engineering, The University of Texas at Austin, TX 78712, USA (e-mail: avinab.saha@utexas.edu, bovik@ece.utexas.edu). Christian Häne, Jean-Charles Bazin, Ioannis Katsavounidis, and Alexandre Chapiro, are with Meta Platforms Inc., Menlo Park, CA 94025, USA (e-mail: chaene@meta.com, jcbazin@meta.com, ikatsavounidis@meta.com, achapiro@meta.com). Yu-Chih Chen is with the Department of Computer Science, National Yang Ming Chiao Tung University, Hsinchu 300093, Taiwan (email: berriechen@nycu.edu.tw). This work was conducted prior to Yu-Chih Chen's employment at NYCU, and she was not supported by any grant.

All experiments, data collection, and processing activities were conducted by the University of Texas at Austin. Meta was involved solely in an advisory role, and no experiments, data collection, or processing activities were conducted on Meta infrastructure.

This paper has supplementary downloadable material available at http://ieeexplore.ieee.org, provided by the author. The material includes a supplemental pdf document. Contact avinab.saha@utexas.edu for further questions about this work.

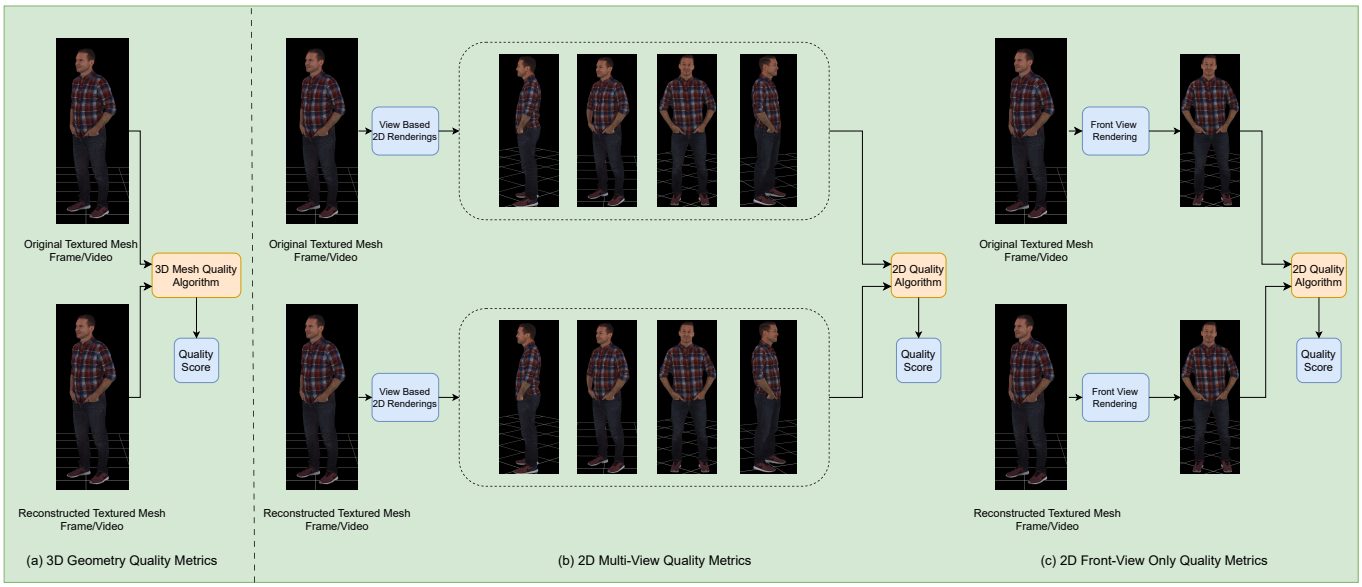

Fig. 1. Illustration of popular approaches to mesh quality assessment using the asset "Frank Casual Talking" from LIVE-Meta Rendered Human Avatar Database. Best viewed Zoomed.

tools are required to ensure network throughput with minimal congestion. Compression causes distortions, however, which can severely degrade viewers' experiences. Other frequent sources of visual artifacts in VR/AR arise from point-of-view-based rendering errors and network-related latency. To avoid these issues, predictions of perceived distortion can be used to in development and monitoring. Consequently, the design of perceptual VQA models for digital humans is important. In modern streaming systems, perceptual models like SSIM [5], [6], VMAF [7] guide encoder and bitrate decisions [8] to achieve optimized visual quality at lower bitrates. These tools help ensure that videos streamed and shared over bandwidth-limited networks maintain the best possible perceptual quality to support pleasurable viewer experiences. Building on these successes in generic video, we believe that a domain-specific perceptual VQA model for rendered digital humans could be used similarly to optimize visual quality and enhance user experiences in virtual telepresence scenarios under VR/AR bandwidth constraints. Designing VQA models and algorithms for VR/AR digital human content requires addressing a number of challenges not encountered in traditional VQA scenarios:

- Portions of the content (e.g. faces, hands) have a high perceptual weight [9], and consequently their visual quality is especially important.
- Immersive VR/AR content typically has a significantly larger field-of-view compared to traditional display. However, the effective resolution in pixels-per-visual-degree is typically much lower [10], which can cause visible loss of quality, or increase the visibility of visual distortions [11].
- User's eye and head motion may interact with display latency and persistence, causing additional artifacts.
- Use of Region-of-Interest (RoI) based encoding, where gaze location is measured (e.g., by eye tracking) or predicted (via saliency analysis) for higher quality encoding.

This work introduces ***Holo****gram* ***Q****uality* ***A****ssessor* (HoloQA), a full reference video quality assessment model for rendered human avatars which uses a mixture-of-experts approach to model low-level visual artifacts and their visual impacts on high-level VR/AR content. HoloQA achieves state-of-the-art quality prediction performance on the newly introduced LIVE-Meta Rendered Human Avatar Database [12]. We also demonstrate the competitive performance of HoloQA on other digital human avatar databases and for synthetically generated video quality applications like cloud gaming.

The paper is organized as follows. Section II reviews prior research on objective video quality assessment algorithms. Section III reviews the LIVE-Meta Rendered Human Avatar Database. Section IV introduces our proposed FR-VQA model for avatar-centric VR/AR and explains the modules of *HoloQA* in detail, while Section V studies the performance against other models. Section VI describes the outcomes of ablation studies on *HoloQA*. Section VII explores potential directions for future research. The Supplementary Material includes more detail on the LIVE-Meta Rendered Human Avatar Database and *HoloQA* model. It also offers a study of *HoloQA's* performance on VQA databases relevant to VR/AR applications.

II. Related Work

We begin by reviewing perceptual objective models for evaluating the quality of 3D meshes and established models for assessing generic 2D image and video quality. The models we consider can all be applied to predict the quality of digital human avatar videos, albeit with varying degrees of success

as this is not part of the design of most metrics. Our target application scenarios are online VR/AR collaboration platforms, such as Meta Horizons. These platforms generally use front-facing cameras, and the LIVE-Meta Rendered Human Avatar Database is designed for creating models that assess quality primarily on the view rendered from the perspective of the front camera.

Perceptual objective quality assessment algorithms encompass two primary categories: Full Reference (FR) and No Reference (NR) models. FR algorithms assess the quality of distorted images, videos, or meshes by comparing them to pristine source content. NR algorithms assess quality by measuring intrinsic properties against statistical models without needing a reference. Our focus here is on the development of FR-VQA models for human avatar videos.

A. 2D Image and Video Quality Assessment

2D FR image and video quality prediction algorithms aim to measure the degree of perceptual fidelity between distorted and reference images and videos. Simple, computationally inexpensive metrics like Peak Signal Noise Ratio (PSNR) and Mean Square Error (MSE) have long been used to measure image and video fidelity; however, they have been shown to poorly align with human perceptions of visual distortion [13]. Early works in perception-based FR-IQA include the Structural Similarity Index (SSIM) [5] and its many variants, including MS-SSIM [6] and FSIM [14]. SSIM measures the perceptual similarity between images by analyzing local luminance, contrast, and structure differences. SSIM is typically extended to video quality measurement by average pooling SSIM scores across the frames of a quality-analyzed video. Visual Information Fidelity (VIF) [15] is another popular FR-IQA model that has been adapted into many FR-VQA models. VIF deploys a perceptually relevant statistical model of bandpass image/video coefficients to represent distortions. FR-VQA models like ST-RRED [16], SpEED-QA [17], and ST-GREED [18] extend the principles introduced in VIF. The Video Multi-Method Assessment Fusion (VMAF) model [7], which, like SSIM, is widely used by industry, is an ensemble model that combines four VIF features with DLM [19], and a temporal frame difference feature. The FUNQUE model proposed in [20], [21] combines many perception-based quality features in a very efficient way by sharing bandpass decomposition to compute them. FovVideoVDP [22] incorporates display device characteristics, bandpass decomposition, and low-level perceptual models such as contrast sensitivity functions and masking, among others.

More FR-IQA/VQA models harness deep learning to predict quality. LPIPS [23] is a popular image similarity metric, often used as an FR-IQA metric, especially in super-resolution problems. DeepVQA [24] uses a CNN-based feature extractor to measure spatiotemporal distortions, amalgamating frame-wise quality scores over time. C3DVQA [25] employs 3D convolutional layers to assess potential temporal aliasing artifacts. Vision Transformers, known for their prowess on diverse visual tasks, have also been utilized for VQA applications. The model in [26] combines CNN-based feature extraction with a Transformer-based encoder to conduct video quality prediction. Self-supervised pre-trained models like CONTRIQUE-FR [27], CONVIQT-FR [28], and Re-IQA-FR [29] trained on large corpora of unlabeled image and video data, demonstrate impressive IQA/VQA capabilities and achieve state-of-the-art performance across diverse subjective quality databases. Recently, large multimodal/language models (LMM/LLMs) [30]–[32] have been applied to image and video quality assessment problems, achieving state-of-the-art results on generic IQA and VQA tasks. However, while powerful models have been developed on large generic datasets, domain-specific tasks such as rendered avatar-quality assessment have been limited by the availability of only small dedicated datasets, obtained by headset-based VR/AR studies requiring carefully controlled conditions and rigorous evaluation procedures. This limited data, coupled with modality and display-specific artifacts (e.g., wide-FOV distortions), has restricted direct transfer of existing method, thereby motivating us to explore tailored approaches, resulting in our model called HoloQA.

B. 3D Mesh Quality Assessment

FR algorithms designed to assess the perceptual quality of 3D meshes that are transmitted over bandwidth limited networks, can be classified into two main categories. One category [33]–[35] directly processes the 3D meshes, while another uses 2D FR-IQA/VQA models to analyze multiple 2D rendered views [36], [37] of 3D models. Fig. 1 shows visual representations of the two primary categories of FR 3D mesh quality assessment models. Fig. 1 (a) depicts direct quality assessment of 3D meshes, whereas (b) illustrates methods involving that render a few views, and then apply 2D-IQA/VQA methods on them.

The authors in [33] used curvature information to objectively assess the quality of 3D meshes. They utilized two psychological aspects of the human visual system, visual masking and saturation effects, to quantify structural changes. The metric in [34] relies on a local roughness measure extracted from Gaussian curvature on the mesh. It calculates a perceptual distance between two meshes by comparing the variance in the normalized surface integrals of the local roughness measure. The method in [35] uses curvature statistics for mesh quality assessment over multiple scales similar to multi-scale SSIM [6]. The authors of [38] propose an SSIM-like method for comparing structural information between an original and a distorted mesh, employing a multi-scale visual saliency map to compute local statistics. As discussed earlier, multiple 2D renderings can be used to simplify mesh quality prediction using a 2D IQA/VQA framework. Methods like those in [39], [40] predict mesh quality by employing a limited set of rendered 2D views.

Given our primary focus on digital human avatars that are transmitted over bandwidth limited networks, it is pertinent to discuss prior studies on assessing the quality of digital human meshes. The DHHQA dataset [41] introduced a large-scale assessment database tailored for building models that evaluate the quality of 3D scanned digital human heads (DHHs). Using this resource, an FR mesh quality assessment model was built,

using 2D projected views and leveraging a pre-trained Swin Transformer [42]. While the DHHQA database contains only 3D human head models, the SJTU-H3D [43] subjective quality database is devoted to full-body digital humans. It comprises 40 high-quality reference digital human bodies along with 1,120 human-labeled versions of them generated by applying seven types of distortions. Another recent resource is the large-scale dynamic digital human quality assessment (DDH-QA) database [44]. The dataset encompasses diverse motion contents and diverse distortions, making possible studies of the perceptual quality of dynamic digital human representations. Geo-metric [45] ran a study to quantify perceived distortions in the 3D geometry of faces, and following FaceMap [9] quantified the saliency of regions of the face when similarly distorted. Finally, the most recent database focused on quality assessment of digital human avatars is the LIVE-Meta Rendered Human Avatar VQA database [12], which we use in the development of our algorithm *HoloQA* design and is discussed in Section III.

C. Relevance and Significance of Our Work

As described in Section II-B, one way to reduce the computational complexity of the 3D mesh quality assessment tasks is to transform them into 2D IQA/VQA problems. Similarly, *HoloQA* operates only on front-rendered views to conduct quality assessment as shown in Fig. 1 (c). Our approach is motivated in two ways: first, the front view of a digital human is usually the primary focus of interest, especially in our target application of VR Live Collaboration. Second, it leads to improved computational efficiency by reducing both the rendering requirements and the processing load on the 3D mesh. *HoloQA* is the first attempt to model the quality of digital human avatar videos within VR viewing environments. Our proposed method includes a number of unique elements designed to meet the needs of VR environments while being flexible enough for traditional 2D VQA. It operates on a wide Field of View (FOV), accounts for the viewing conditions in an HMD, and is sensitive to temporal distortions associated with the natural head movements of users wearing VR headsets. Additionally, we perform region-of-interest-specific processing along with semantic information processing, which in our context involves analysis of the human body and face. As a result, *HoloQA* achieves significantly better performance than existing methods in assessing the performance of digital human rendering.

III. LIVE Meta Rendered Hologram VQA Database

This section presents a brief review of the recently introduced LIVE-Meta Rendered Human Avatar VQA Database [12]. This subjective video quality database targets VR Live Collaboration applications, and contains 720 video sequences at 1800×1920 resolution, with durations between 14-15 seconds. Videos were obtained by introducing artifacts onto 36 pristine source avatar videos using 20 different distortion parameter settings. The authors conducted a large-scale subjective study, with 78 human subjects rating each video. All videos in the study consisted of front view renders of 3D human body avatars. They were viewed by the study participants in a VR headset (Meta Quest Pro [46]). Sample frames from this dataset are shown in Fig. 2. The distortions studied include latency, low frame rate, decreased texture map resolution, and reduced depth map resolution (see Section I of the Supplementary Material for more details). The diversity of content and distortions in the database make it suitable for developing quality prediction models for digital human applications.

IV. Proposed Algorithm: *HoloQA*

The *Mixture of Experts* (MoE) concept is a popular design principle in machine learning. It emphasizes the integration of multiple specialized "experts" to handle complex tasks by leveraging the combined input of multiple component models. The NR-IQA model Re-IQA [29] successfully deployed the concept of MoE using two separate encoders that respectively analyze high-level image content and low-level visual distortions. In a similar manner, *HoloQA* adopts an MoE approach that deploys two separate modules predictive of perceptual quality-related aspects of avatar distortion and content. Unlike Re-IQA, which relies on an extensive dataset of unlabeled natural scene images to pre-train a quality-aware encoder, we created a handcrafted module for two reasons. First, there is a lack of available 2D-rendered video frames portraying digital human avatars, which is inadequate for pre-training deep neural network backbones in self-supervised settings. Second, we integrate essential feature extraction mechanisms into the distortion-aware module in HoloQA, imbuing it with characteristics unique to VR-related applications that conventional convolutional neural networks/transformers cannot readily capture in the absence of large amounts of data. The content-aware module of *HoloQA* builds on the same vanilla ImageNet [47] pre-trained backbone used in Re-IQA, by further pre-training on avatar content-specific datasets, such as human face and body image databases. A schematic of *HoloQA* is shown in Fig. 3.

A. Distortion-Aware Module

The perceptual distortion-aware module of HoloQA has two primary components. The first is the Human Visual System (HVS) simulator, which embodies various perceptual processing models, including multiscale decomposition, contrast encoding, and contrast sensitivity function (CSF). The second component, a statistical feature extractor, processes the outputs generated by the HVS simulator. It measures the distortion-sensitive statistical differences between distorted videos and their pristine reference counterparts, yielding a set of avatar video distortion-aware features.

1) HVS Simulator: The HVS Simulator comprises four sequential processing steps, accepting input videos and generating activation maps fed to the statistical feature extraction module. The sub-modules in the HVS Simulator were inspired by models of displays and human vision popularized in the VDP metrics [22], and suitably adapted to our use case. Below, we reproduce and cite the source of

(a) Amanda Seated Listening (b) Luke Seated Listening (c) Natasha Serious Talking (d) Wendy Listening Business

Fig. 2. Sample Frames with names of video sequences from the 2D rendered videos in LIVE-Meta Rendered Human Avatar Database.

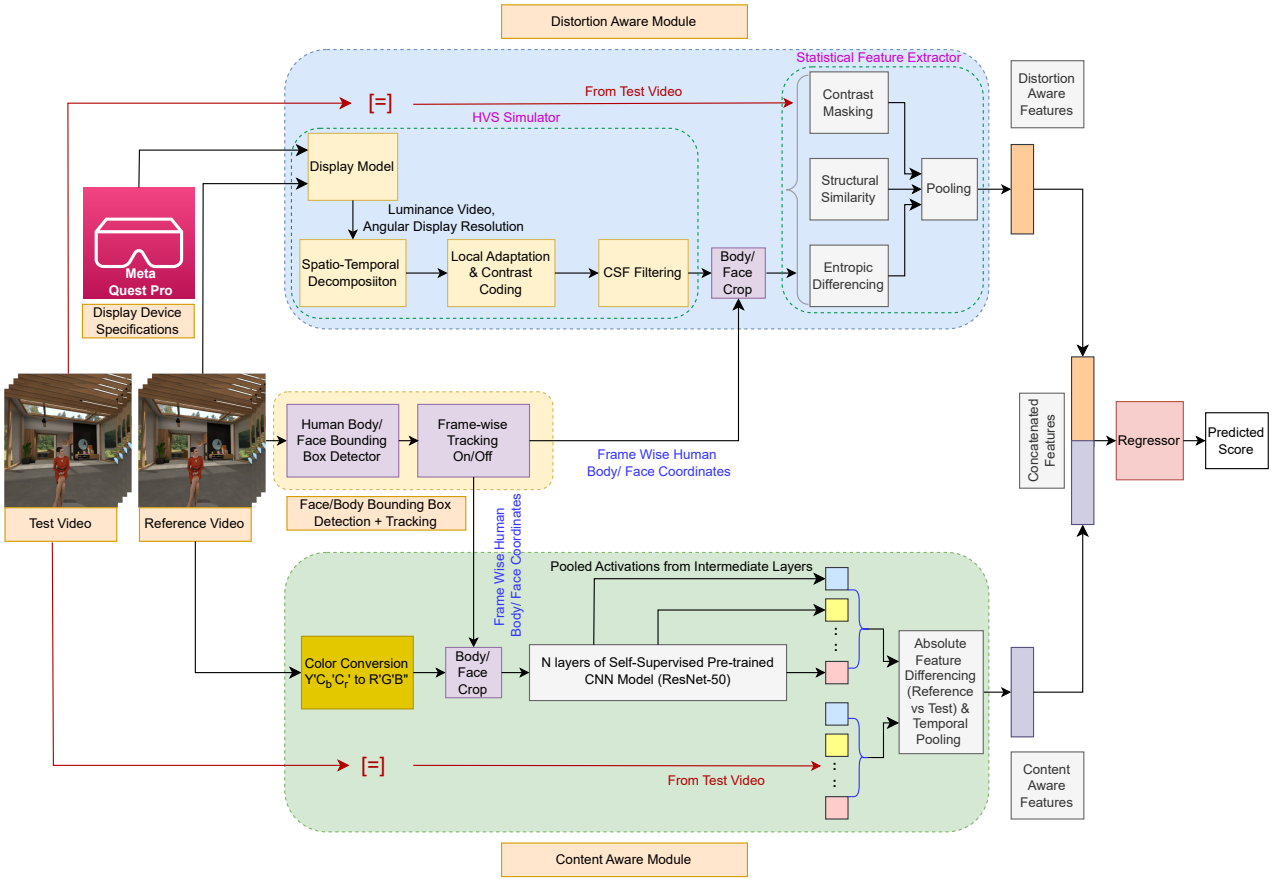

Fig. 3. Schematic flow of the HoloQA model. Best viewed zoomed.

each component, as appropriate. Our display model targeted the Meta Quest Pro (specifications given in Table I). Notably, this can be modified to accommodate other display devices and VQA applications, as demonstrated in the Supplementary.

a. Display Model: Modeling of the display characteristics begins with display photometry. This component is responsible for converting encoded pixel values to linear physical units (cd/m^2), accounting for the display characteristics (see Section IIA of the Supplementary material for more detail). During the development of *HoloQA*, videos are stored in an 8-bit Y'Cb'Cr format, decoded with *ffmpeg* and converted into floating-point representations. They are then transformed into the R'G'B' format (encoded sRGB). Following this, the encoded sRGB colors are converted into relative linear units of luminance :

$$srgb2linear(p) = \begin{cases} \left(\frac{p+0.055}{1.055}\right)^{2.4} & \text{if } p > 0.04045 \\ \frac{p}{12.92} & \text{otherwise} \end{cases}, \quad (1)$$

where p is an encoded pixel value in the R'G'B' format in the range [0,1]. Finally, the decoded RGB values are scaled

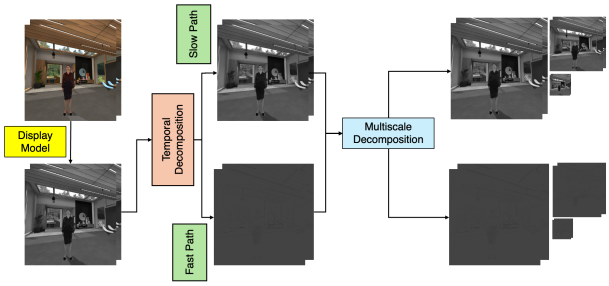

Fig. 4. Block diagram of spatio-temporal decomposition used in the Distortion Aware Module of HoloQA. Best viewed electronically, zoomed.

to obtain absolute units of light :

$$L = (Y_{peak} - Y_{black}) \times srgb2linear(p) + Y_{black}, \quad (2)$$

where Y_{peak} is the maximum luminance, and Y_{black} is the minimum luminance of the Meta Quest Pro display (Table I). It should be noted that Y_{black} corresponds to the minimum luminance of the screen only when the ambient luminance measured on the display screen is close to zero, an assumption valid for the VR display device we employ. This process transforms the input visual signal into a luminance signal for subsequent processing by the other blocks in the HVS Simulator. The display photometry model presented here is a simplified adaptation of the perceptual display model proposed in [48].

After display photometry, the geometric properties of the display are modeled, calculating the angular display resolution as described by [22]. In traditional VQA studies, the displays may cover only a small field of view, and the angular resolution measured in pixels-per-visual-degree (ppd) is assumed to be constant across the display and can be obtained using the relation:

$$n_{ppd\_center} = \frac{\pi}{360 \times tan^{-1}\left(\frac{0.5 \times d_{\text{width}}}{r_h \times d_v}\right)}, \quad (3)$$

where d_{width} is the width of the width of the virtual VR display, d_v is the viewing distance in the VR environment, and r_h is the horizontal resolution of the display device. From Table I, d_v and r_h are readily available, but to compute n_{ppd_center}, we need d_{width}, which needs to be determined by a series of computations, as shown in Section IIB of the Supplementary material. However, for wide FoV VR displays like the Meta Quest Pro, the angular resolution is not constant and increases considerably as the deviation from the central viewing direction increases. The following relation can be used to compute the angular resolution as a function of eccentricity e (in degrees) :

$$n_{ppd}(e) = n_{ppd\_center} \times \frac{\tan\left(\frac{\pi e}{180} + \frac{1}{2 \cdot n_{ppd\_center}}\right) - \tan\left(\frac{\pi e}{180}\right)}{\tan\left(\frac{1}{2 \cdot n_{ppd\_center}}\right)}. \quad (4)$$

In Section IIC of the Supplementary material, we plot the variation of angular resolution as a function of eccentricity.

TABLE I
META QUEST PRO SPECIFICATIONS

Resolution	FoV (° diagonal)	Viewing Distance (m)	Max Luminance (nits)	Min Luminance (nits)
1800×1920	111.24	1.2	200	0.1

b. Spatio-Temporal Decomposition: The temporal decomposition employed *HoloQA* follows the model of [22], which in turn draws inspiration from prior research on retinal ganglion cells in the vision system [49]–[52]. Approximately 80% of these cells are Parvocellular (P-cells), while 15-20% are Magnocellular (M-cells). M-cells demonstrate heightened responsiveness to rapid temporal changes but exhibit less sensitivity to spatial detail or color. Conversely, P-cells are highly responsive to fine spatial detail and color information but are less so to temporal stimuli. As shown in Fig. 4, the spatio-temporal decomposition has dual paths, with one pathway capturing rapid changes and the other capturing slower ones. In [22], the time domain response of the paths encoding slow changes is expressed as by an exponential function of logarithmic time t:

$$R_{Slow}(t) = k_1 \exp\left(\frac{(\log(t+\epsilon) - \log(\beta_s))^2}{2\sigma_s^2}\right), \quad (5)$$

where ϵ is a small constant that ensures the stability of the logarithmic function near $t = 0$, β_s is the time lag of the response, and σ_s determines the filter's bandwidth. The constant k_1 is chosen so that the filter's response to a constant signal remains unchanged, i.e., is level preserving. The time domain response of the path that encodes fast changes is obtained as the derivative of the time domain response of the slow pathway :

$$\begin{aligned} R_{Fast}(t) &= k_2 \frac{d}{dt} R_{Slow}(t) \\ &= -k_2 \frac{R_{Slow}(t)\,(\log(t+\epsilon) - \log(\beta_s))}{\sigma_s^2 (t+\epsilon)} \end{aligned} \quad (6)$$

As in [22], the normalization constant k_2 is chosen to ensure that the contrast at the peak frequency of 5 Hz is preserved. The peak responses of the static and transient paths are at 0 Hz and 5 Hz, respectively. The two paths are linear digital filters applied as sliding windows on the output of the display model. Next, similar to popular FR-IQA/VQA algorithms [5], [6], [22] that employ spatial decompositions to model the primary visual cortex, the responses from the slow and fast encoding paths are subjected to spatial decompositions. While older models such as [5], [6] utilize simple spatial multi-scaling, later models like [15], [17], [18] employ wavelet-based decompositions, which have better approximation properties but higher computational complexity. As in [22], we employ a decimated Laplacian pyramid-based decomposition [53] applied independently on the slow and fast paths. The decimated Laplacian pyramid is computationally efficient as compared to wavelet decompositions. Section IID of Supplementary Material provides a high-level understanding of how to construct a decimated Laplacian pyramid. The peak

frequency in each band indexed by b is given by :

$$\rho_b = \begin{cases} 0.5\, n_{ppd} & \text{for } b = 1 \\ \frac{0.1614}{2^{b-2}}\, n_{ppd} & \text{for } b > 1 \end{cases}, \tag{7}$$

where n_{ppd} is the angular resolution given by equation 4. Unlike [22], which dynamically adjusts the depth of the Laplacian pyramid based on the characteristics of the display geometry, we limit its depth to three levels, as this was found not to reduce the model's performance (across multiple VQA datasets), but significantly reduces the time complexity to obtain distortion-aware features, an observation also verified in [54]. Fig. 4, shows temporal and spatial decompositions on a video response obtained from the display model.

c. Local Adaptation and Contrast Coding: Following the temporal and spatial decomposition of the visual signal into two temporal channels and three pyramid levels, we compute the contrast at each coordinate of the pyramid coefficients. Let $L_{b,c}(x, y)$ and $G_{b,c}(x, y)$ denote the Laplacian and Gaussian pyramid coefficients at pixel locations (*x,y*), where $b \in [1, 3]$ refers to the pyramid levels and $c \in \{slow, fast\}$ refers to the temporal channels. Contrast (or Michelson's contrast) C can be expressed as :

$$C = \frac{L_{max} - L_{min}}{L_{max} + L_{min}} = \frac{\Delta L}{L_{mean}}, \tag{8}$$

where L denotes luminance. We use the Laplacian and Gaussian pyramid coefficients to compute local bandpass and smoothed contrast. Since any bandpass signal can be regarded as the difference between two low-pass (in this case, Gaussian-smoothed) signals, the Laplacian coefficients, respectively, can be used as smoothed proxies for ΔL. As observed in [55], the contrast thresholds are better aligned with Weber's Law when L_mean in equation 8 is replaced by background luminance L_a, which only depends on local neighborhoods of luminance. We approximate luminance at each coordinate and each pyramid level by the smoothed luminance values at one higher level in the corresponding Gaussian pyramid of the slow encoding pathway. Thus, contrast $C_{b,c}(x, y)$ is computed as :

$$C_{b,c}(x, y) = \frac{L_{b,c}(x, y)}{G_{b+1,Slow}(x, y)}. \tag{9}$$

d. Spatio-Temporal CSF Filtering: Each computed contrast is normalized by converting it into multiples of the threshold contrast. This is done by multiplying the contrast with the sensitivity, defined as the inverse of threshold contrast. The sensitivity is obtained from the contrast sensitivity function (CSF). The Let $C'_{b,c}$ denote normalized contrast and $S_{b,c}$ denote the CSF. The relation among $S_{b,c}$, $C_{b,c}$ and $C'_{b,c}$ at (x, y) is given by :

$$C'_{b,c}(x, y) = C_{b,c}(x, y) \cdot S_{b,c}(x, y). \tag{10}$$

We utilize the CSF model developed in [22], which incorporates extra-foveal correction applied to the standard spatiotemporal CSF, which better matches the wide field-of-view visualized in VR applications. It is given by :

$$S_{b,c}(x, y) = S_{exfov}(\rho_b(x, y), \omega_c, L_a(x, y), e(x, y)). \tag{11}$$

The extra-foveal CSF is a function of the peak spatial frequency in each band $\rho_b(x, y)$ at coordinates (x, y), ω_c is the peak temporal frequency of the slow/fast encoding path, $L_a(x, y), e(x, y)$ is the local luminance of adaptation and the eccentricity at the coordinates (x, y), assuming that the fixation point is at the center of the frame. More details about the CSF can be found in Section IIE of the Supplementary section.

2) Statistical Feature Extractor: Let the response of the HVS simulator be $R_{(b,c)}(x, y)$. The final step of the HVS simulator involved normalizing the contrast using the CSF function, $R_{(b,c)}(x, y) = C'_{b,c}(x, y)$. We then apply perceptually relevant statistical feature extractors to convert contrast responses into "distortion-aware" features. The responses of the HVS simulator are transformed into distortion-aware feature maps, which are first spatio-temporally pooled to generate distortion-aware atomic features. These features are then used to train a regression model to predict perceptual quality.

A variety of perceptually motivated statistical feature extractors are available, including structural similarity [5], [6], [56], entropy differencing [16]–[18], and contrast masking [19], [22]. The feature extraction process for all these families of feature extractors is relatively inexpensive as compared to the HVS simulation pipeline. Consequently, multiple feature extractors can be applied without significantly increasing the overall computational complexity of the model [7], [20]. In the following, we describe the statistical feature extraction process used by the *HoloQA* model. Throughout, denote the responses of the HVS simulator to the reference and test videos as $R^{ref}_{(b,c)}(x, y)$ and $R^{test}_{(b,c)}(x, y)$, videos respectively.

a. Structural Similarity: Inspired by the success of CW-SSIM [56] on quantifying image similarity, we employ a similar model on the responses $R^{ref}_{(b,c)}(x, y)$ and $R^{test}_{(b,c)}(x, y)$ delivered by the HVS simulator. However, while CW-SSIM computes image similarity in the complex wavelet domain, the Laplacian pyramid yields real-valued responses. The structural similarity map is thus defined as,

$$SS_{(b,c)}(x, y) = \frac{2 \cdot G * (R^{ref}_{(b,c)}(x, y) \cdot R^{test}_{(b,c)}(x, y)) + \epsilon_1}{G * ((R^{ref}_{(b,c)}(x, y))^2 + (R^{test}_{(b,c)}(x, y))^2) + \epsilon_2}, \tag{12}$$

where ϵ_1 and ϵ_2 are small stabilizing constants ($\approx 10^{-12}$) and $G(.)$ is a 7×7 box filter that smooths the activations.

b. Entropic Differencing: Several popular VQA models are based on measuring entropic differences (ED) between bandpass distorted videos and their bandpass reference counterparts subjected to identical spatial and/or temporal decompositions [17], [18]. We have found (as shown in Section IIG of Supplementary Material) that the bandpass responses to the preprocessed reference videos $R^{ref}_{(b,c)}(x, y)$ reliably obey a zero-mean generalized gaussian distribution (GGD), while the bandpass responses to the preprocessed distorted videos $R^{test}_{(b,c)}(x, y)$ deviate from GGD. This implies that simple measurements that quantify those deviations from the GGD, e.g., by entropic differencing, can be used to generate highly

predictive distortion-aware features. The univariate probability density function (PDF) of a zero mean GGD variate is given by:

$$f_X(x) = \frac{\beta}{2\alpha\Gamma(1/\beta)} \exp\left(-\left(\frac{|x|}{\alpha}\right)^{\beta}\right), \tag{13}$$

where $\Gamma(\cdot)$ is the Gamma function,

$$\Gamma(a) = \int_0^{\infty} x^{a-1} e^{-x}\, dx. \tag{14}$$

The shape parameter β controls the shape of the distribution (tail weight and peakiness) while α scales the variance. The closed-form expression for the entropy of a zero mean GGD is:

$$h(X) = \frac{1}{\beta} - \log\left(\frac{\beta}{2\alpha\Gamma(1/\beta)}\right). \tag{15}$$

The derivation of the closed-form expression for entropy for a zero-mean GGD can be found in Section IIF of the Supplementary material. To compute these quantities, the responses $R^{ref}_{(b,c)}(x,y)$ and $R^{test}_{(b,c)}(x,y)$ are divided into 5×5 spatial blocks indexed as $R^{ref}_{(p,b,c)}(x,y)$ and $R^{test}_{(p,b,c)}(x,y)$ where p $\in [0, P-1]$. On each of the local spatial blocks, compute two terms, local variance $\sigma^2(R^{ref/test}_{(p,b,c)})$ and the local entropy, $h(R^{ref/test}_{(p,b,c)})$. Similar to [16], [18], we also scale the entropy using the scaling parameter $\gamma(R^{ref/test}_{(p,b,c)})$. The parameter is defined as:

$$\gamma(R^{ref/test}_{(p,b,c)}) = log(1 + \sigma^2(R^{ref/test}_{(p,b,c)})). \tag{16}$$

Scaling by the variances imparts greater weight to the entropic differences in spatial/temporal regions of high activity, which tend to be more salient, while providing numerical stability in regions having little activity. Finally, the entropic differences after scaling are obtained as:

$$ED_{(p,b,c)}(x,y) = |\gamma(R^{ref}_{(p,b,c)})h(R^{ref}_{(p,b,c)}) - \gamma(R^{test}_{(p,b,c)})h(R^{test}_{(p,b,c)})|. \tag{17}$$

The scaled entropies across all patches are mapped back to their original spatial locations, yielding a final entropic differencing map $E_{(b,c)}(x',y')$, where $x' = \lfloor \frac{x}{5} \rfloor$, and $y' = \lfloor \frac{y}{5} \rfloor$.

c. Contrast Masking: a phenomenon in visual perception where the visibility of an image feature is reduced due to the presence of another feature, usually having similar spatial frequencies and/or orientations. Masking effects play a crucial role in video processing applications, including video compression and quality assessment, since successful masking models can be used to predict how humans perceive distortions in complex scenes. In VQA algorithms, the reference video usually serves as the masking signal to adjust distortions, as distortions are commonly less visible on textured areas than on smoother ones. However, some artifacts render a video more textured or locally active than the original, making it problematic to use the original video as the masker; a good example is the introduction of the false contours or "banding" of smooth areas, often from compression quantization. One way of handling this is via mutual masking, first introduced in MOVIE [57], which can be modeled as :

$$MM_{b,c}(x,y) = \min\left\{\left|R^{\text{ref}}_{b,c}(x,y)\right|, \left|R^{\text{test}}_{b,c}(x,y)\right|\right\}. \tag{18}$$

Then, the output map of the contrast masking operator becomes :

$$CM_{b,c}(x,y) = \frac{\left|R^{ref}_{b,c}(x,y) - R^{test}_{b,c}(x,y)\right|^p}{1 + (k \times MM_{b,c}(x,y))^{q_c}}, \tag{19}$$

where *k*, *p* and q_c, $c \in \{slow, fast\}$ are free parameters of the model. We adopted the parameter settings from [22] to avoid additional tuning.

d. Atomic Feature Generation: Next, we outline the steps to develop atomic features from the three statistical space-time feature maps $SS_{(b,c)}(x,y,t)$, $ED_{(b,c)}(x,y,t)$, and $CM_{(b,c)}(x,y,t)$, where we have now added the frame index t. In the following, we describe the process of pooling the statistical feature maps along their spatial dimensions to derive atomic features. We will explain the process using $SS_{(b,c)}(x,y)$, but the process is same of the other maps $ED_{(b,c)}(x,y)$, $CM_{(b,c)}(x,y)$. At each time frame instant *t*, compute the Coefficient of Variation (CoV) pooling :

$$SS_{CoV}(t) = \frac{std(SS_{(b,c)}(x,y,t))}{mean(SS_{(b,c)}(x,y,t))}, \tag{20}$$

where *mean* and *std* are the average and standard deviation calculated over x, y. Executing the identical process (20) on the other maps yeilds $ED_{CoV}(t)$ and $CM_{CoV}(t)$. Next, mean and standard deviation pooling of SS_{CoV}, ED_{CoV}, and CM_{CoV} across the time dimension yeilds distortion-aware atomic features of the form $mean(SS_{CoV}(t))$ and $std(SS_{CoV}(t))$, and similar for ED_{CoV}, and CM_{CoV} (hence, six in total). Mean and standard deviation pooling capture average and average deviations of the distortion-aware maps. Large values of the latter are expressive of large quality variation, which leads to worse impressions of overall quality. This is effectively captured by CoV, as explained in [57]. Two atomic features are computed at each level/band of the distortion-aware feature maps. Since we use 2 temporal channels $\times$ 3 spatial decomposition levels, a total of 12 atomic features are generated on each three classes of distortion-aware feature map, hence 36 atomic features are regressed on when learning the Distortion-Aware module of *HoloQA*.

B. Content-Aware Module

Contemporary VQA algorithms [29], [58], [59] commonly use pre-trained neural networks to model those aspects of content on quality prediction tasks. In a generic VQA scenario, where videos typically feature diverse content, ImageNet pre-trained models have been shown to improve quality prediction when combined with distortion-aware features [29], [58]. The more sophisticated Distortion-Aware Module in HoloQA delivers superior or comparable performance to other state-of-the-art VQA algorithms, as evidenced by the results in Tables II and VII, without relying on content-specific features or pre-training on other unlabeled image or

video databases. The content-aware module further enhances quality prediction performance by learning quality-relevant aspects of identifiable content, such as faces, human bodies, or parts of bodies. A content-aware module can thus strongly contribute to more accurate overall video quality evaluations.

1) Content Specific Self-Supervised Fine-Tuning: We used the ResNet-50 based MoCo-v3 [60] model, pre-trained on the ImageNet database, as the backbone of the content-aware module. For our use case of quality assessment of digital human avatars, we created two content-specific models-one for the human body and one for the human face-by fine-tuning the pre-trained MoCo-v3 model in a self-supervised setting. As explained below, we fine-tuned models on both synthetic human faces and on full-length human bodies, presupposing these to be the most visually salient regions of live collaboration scenarios. Sections IV-C and IV-D provide details on the RoI processing.

a) Fine-Tuning Databases: We used two databases of synthetically generated images to fine-tune the ImageNet-pre-trained MoCo-v3 model. Human body images were sourced from the SHHQ database [61], while human face images were obtained from the VGGFace2-HQ database [62]. Additional information about these datasets can be found in Section IIIA of the Supplementary material.

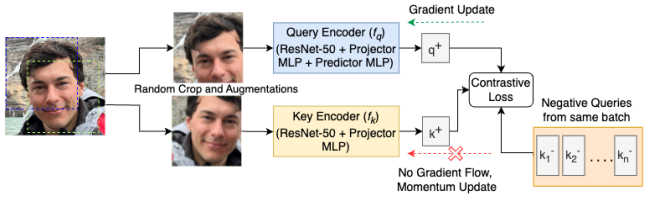

Fig. 5. Block diagram of MoCo-v3 based content aware finetuning workflow. This example shows the processing of face images; the workflow remains same when fine-tuning on human body images. Best viewed zoomed.

b) Fine-Tuning Algorithm: We fine-tuned the ImageNet pre-trained MoCo-v3 models on the two above-mentioned content-specific databases. The setup, as shown in Fig. 5, consists of two encoders, f_q and f_k. The encoder f_q comprises a ResNet-50 backbone, a projection head, and an additional prediction head as in [63]. The projection and prediction heads are feed-forward multi-layer perceptrons. By contrast, the encoder f_k includes the backbone and projection head but lacks the prediction head. The f_k encoder is updated by taking the moving average of f_q [64], without the prediction head.

During fine-tuning, we generate two augmented crops of every face (or body) image, reusing the default data augmentation techniques used in MoCo-v3: random resized cropping, horizontal flipping, color jittering, grayscale conversion, blurring, and solarization. The two face (or body) image crops are processed by two separate encoders, f_q and f_k, resulting in output vectors q and k. From, Fig. 5, q acts as a query to retrieve its corresponding key k. This is accomplished by minimizing the contrastive loss, InfoNCE [65], ensuring that the query and key vectors are optimally aligned. The loss for query q can be expressed as :

$$L_q = -\log \frac{\exp\left(q \cdot k^+/\tau\right)}{\exp\left(q \cdot k^+/\tau\right) + \sum_{k^-} \exp\left(q \cdot k^-/\tau\right)} \tag{21}$$

where k^+ is the output of f_k for the crop obtained from the same image as q, which is referred to a positive sample. The set $\{k^-\}$ includes the outputs of f_k from different images, referred to as negative samples with respect to q. τ is a temperature hyper-parameter used for l_2 normalization of q and k. Further details on the fine-tuning process can be found in Section IIIB of the Supplementary material.

2) Atomic Feature Generation The atomic feature generation process in the Content-Aware Module is similar to that of the Distortion-Aware module. Unlike the majority of popular IQA/VQA methods [27], [29], [58], [59], that extract features from the final layers of a pre-trained deep neural networks, we draw inspiration from prior studies that exploit intermediate-layer activations (e.g., [23], [66]) and extract features from intermediate layers. Details on selecting these layers can be found in Section IIIB of the Supplementary Material. For each selected intermediate layer, we obtain a feature map $FM_{(l,c)}(x, y, t)$, where l and c denote the layer index and channel index, respectively, *(x,y)* are spatial indices, and *t* is a frame index. We then apply average and standard deviation pooling across the spatial dimensions to obtain frame-level features, $f_{(l,c)}(t)$ and $g_{(l,c)}(t)$ given by,

$$\begin{aligned} f_{(l,c)}(t) &= \text{mean}_{x,y}(FM_{(l,c)}(x, y, t)) \\ g_{(l,c)}(t) &= \text{std}_{x,y}(FM_{(l,c)}(x, y, t)) \end{aligned} \tag{22}$$

Before applying temporal pooling, we compare the reference and test sequences in feature space. At each layer l and channel c, first obtain frame-level features on the reference $f^{\text{ref}}_{l,c}(t), g^{\text{ref}}_{l,c}(t)$ and the test $f^{\text{test}}_{l,c}(t), g^{\text{test}}_{l,c}(t)$ videos as in equation 22. We then compute absolute differences as

$$\Delta f_{l,c}(t) = \left|f^{\text{ref}}_{l,c}(t) - f^{\text{test}}_{l,c}(t)\right|, \quad \Delta g_{l,c}(t) = \left|g^{\text{ref}}_{l,c}(t) - g^{\text{test}}_{l,c}(t)\right|. \tag{23}$$

Then, to obtain video-level features, we perform mean and standard deviation pooling across the temporal dimension using $\Delta f_{(l,c)}(t)$ and $\Delta g_{(l,c)}(t)$, resulting in four final features at each layer. Extracting features from four intermediate layers yields $4 \times 4 = 16$ features. To reduce the computational complexity, the feature extraction is performed at one frame/sec. We conducted experiments comparing feature extraction using all frames against subsampling feature extraction over discrete time steps (e.g., 1 frame per second, see Section V of the Supplementary material.) and observed negligible differences in performance.

C. Region of Interest-based Processing

The video frames in the LIVE-Meta Rendered Human Avatar VQA database consist of constant, static backgrounds with the digital human avatars overlaid. In this database, the applied distortions only affect the human avatars while

preserving the background regions. To ensure quality prediction focused on the rendered avatars while mitigating interference from the background, we define RoI bounding boxes, similar to models like [67].

Here, we focus on quality assessment, focusing on the full bodies and on the faces of the digital human avatars by computing bounding boxes on these regions. To obtain bounding boxes of body regions, we processed each frame from every reference video in the database, utilizing YOLO-v7 [68]. To compute the bounding boxes of faces, we used the YOLO-v8 face model [69]. By using these high-performance detectors, HoloQA is able to accurately identify the body and facial regions within every video frame of the LIVE-Meta Rendered Human Avatar VQA database. Fig. 6 shows examples of body and face bounding box detection. Additionally, we provide additional robustness analysis of these detectors in Section VII of the Supplementary Material.

D. Model Versions

We developed four versions of *HoloQA* by varying on the way bounding box detection of human bodies and faces is processed, by introducing frame tracking and the backbone used in the content-aware module.

a. HoloQA: In the simplest version of the model, the human body and face bounding boxes are computed on all frames of each analyzed video. Among these, the largest bounding box across all frames is found, whether human body or face. Assuming the video frames have spatial indices increasing from the left to right and from top to bottom, let the bounding coordinate of the face or body bounding box in the i-th frame be denoted as:

$$(top_i, \quad left_i, \quad bottom_i, \quad right_i).$$

Then, across all frames i, the largest bounding box is defined by the bounding coordinates:

$$\left(\min_i left_i, \quad \min_i top_i, \quad \max_i bottom_i, \quad \max_i right_i\right).$$

This ensures that the final region of interest over which avatar video quality is predicted (whether of the body or the face) is consistently contained within a global cropped window. We use this admittedly simple version of *HoloQA* to facilitate comparisons against other FR-VQA algorithms. This is important since nearly all widely used FR-VQA algorithms predict quality over fixed rectangular frames. The backbone used in the content-aware module in this version is the MoCo-v3 model pre-trained on the ImageNet database, enabling its use on non-avatar related VQA applications.

b. HoloQA+: This version aims to improve the performance of the Content-Aware Module by fine-tuning the ImageNet pre-trained backbone in a self-supervised setting using unlabelled human face and body images. As discussed above, while the simplicity of the *HoloQA* design yields a generic algorithm usable for any VQA task, *HoloQA+* is enhanced for predicting the quality of 2D-rendered avatar videos in VR presentations involving human interactions.

c/d. HoloQA/HoloQA+ with Frame Tracking: To further enhance the performance of *HoloQA/HoloQA+*, we deployed a bounding box tracking mechanism to better localize the body and face regions, rather than just using a single, global maximum bounding box. The tracking is achieved by acquiring the bounding boxes and directly utilizing their coordinates in the HoloQA processing pipeline, instead of obtaining the maximal global bounding box as in the previous model versions. This approach improved the algorithm performance of HoloQA/HoloQA+, as shown in Section V.

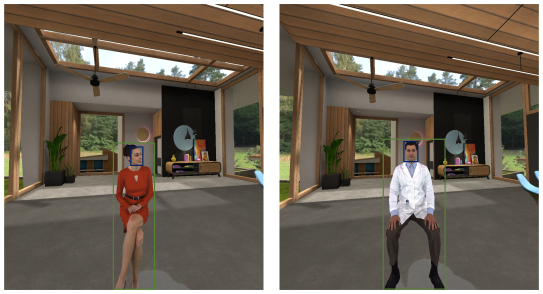

Fig. 6. Sample frames from the sequences "Amanda Seated Listening Party" and "Doctor Luke Seating Listening" showing the overlaid body bounding box (in green) and face bounding box (in blue). Best viewed zoomed.

E. Full Reference VQA Regression

The final predictions produced by (all versions of) *HoloQA* are obtained by concatenating the feature representations produced by the Distortion-Aware and Content-Aware Modules into a single vector. This feature vector is used to train a Support Vector Regressor (SVR) with the radial basis function kernel, to map the features to the subjective video in the LIVE-Meta Rendered Human Avatar database.

V. Algorithm Comparisons

A. Evaluation Protocol

We assessed the effectiveness of all four versions of *HoloQA* against other widely-used FR VQA algorithms using standard performance criteria: Spearman's Rank Order Correlation Coefficient (SROCC), Kendall Rank Correlation Coefficient (KRCC), Pearson's Linear Correlation Coefficient (PLCC), and Root Mean Square Error (RMSE). SROCC and KRCC assess the monotonic relationship between the objective model's predictions and human scores, while PLCC and RMSE gauge prediction accuracy. Before calculating the PLCC and RMSE outcomes, the predicted quality scores of each model underwent a logistic non-linearity transformation [70] to enhance linearity in the objective predictions and align them with the Difference Mean Opinion Score (DMOS) scale. More details on these four performance criteria can be found in Section IV of the Supplementary material.

Each algorithm underwent testing on 1000 random train-test splits of the LIVE-Meta Rendered Human Avatar database using the four performance metrics. In each split, videos

TABLE II
MEDIAN SRCC, KRCC, PLCC, AND RMSE OF FULL-REFERENCE QUALITY ALGORITHMS AGAINST HUMAN JUDGMENTS ON THE LIVE-META RENDERED HUMAN AVATAR VQA DATABASE (1000 TRAIN/TEST SPLITS). ALL ALGORITHMS USE STANDARD GLOBAL BOUNDING BOXES ACROSS TIME STEPS. TOP-3 ALGORITHMS IN EACH BOUNDING BOX CATEGORY ARE BOLDFACED. ENHANCED VERSIONS OF *HoloQA* ARE MARKED WITH 🔥.

Full Reference Algorithm		Regressor	Bounding Box: Body				Bounding Box: Face			
			SRCC (↑)	KRCC (↑)	PLCC (↑)	RMSE (↓)	SRCC (↑)	KRCC (↑)	PLCC (↑)	RMSE (↓)
IQA methods using handcrafted features	PSNR-RGB	N/A	0.7219	0.5223	0.7562	0.5750	0.7129	0.5111	0.7405	0.5924
	PSNR-Y	N/A	0.7167	0.5192	0.7505	0.5798	0.6983	0.4987	0.7285	0.6026
	SSIM [5]	N/A	0.7694	0.5618	0.7755	0.5537	0.8001	0.5929	0.8264	0.4917
	DLM [19]	N/A	0.8599	0.6601	0.8919	0.3995	0.8716	0.6762	0.9106	0.3624
	VIF [15]	N/A	0.7184	0.5200	0.7499	0.5819	0.8154	0.6085	0.8513	0.4621
IQA methods using supervised pre-trained deep features	LPIPS [23] (AlexNet)	SVR	0.8616	0.6573	0.8815	0.4131	0.8790	0.6835	0.9069	0.3703
	LPIPS (VGG)	SVR	0.8441	0.6356	0.8535	0.4558	0.8764	0.6803	0.9047	0.3734
IQA methods using self-supervised pre-trained deep features	CONTRIQUE [27] (Full Reference)	SVR	0.9024	0.7205	0.9470	0.2816	**0.8991**	0.7138	0.9403	**0.2976**
	Re-IQA [29] (Full Reference)	SVR	**0.9068**	**0.7298**	0.9401	0.2919	0.8988	**0.7154**	0.9308	0.3218
VQA methods using handcrafted features	VMAF [7] (v0.6.1)	SVR	0.8179	0.6085	0.8367	0.4789	0.8695	0.6716	0.8949	0.3905
	VMAF (Retrained)	SVR	0.8432	0.6418	0.8732	0.4398	0.8914	0.7022	0.9055	0.3674
	FovVideoVDP v1.2 [22]	N/A	0.7763	0.5750	0.8103	0.5168	0.8810	0.6888	0.9155	0.3527
	ST-GREED [18]	SVR	0.8543	0.6619	0.8800	0.4162	0.8946	**0.7154**	0.9148	0.3552
	SpEED-QA [17]	N/A	0.7415	0.5363	0.7772	0.5493	0.8781	0.6853	0.9103	0.3636
	FUNQUE [20]	SVR	0.8655	0.6684	0.9049	0.3740	0.8980	0.7140	**0.9404**	0.2982
	Y-FUNQUE+ [21]	SVR	0.8663	0.6641	0.8914	0.3971	0.8956	0.7120	0.9355	0.3094
	3C-FUNQUE+ [21]	SVR	0.8599	0.6547	0.8832	0.4104	0.8981	0.7129	0.9357	0.3101
VQA methods using self-supervised pre-trained deep features	CONVIQT [28] (Full Reference)	SVR	0.9065	0.7289	**0.9526**	**0.2673**	0.8834	0.6903	0.9273	0.3283
Handcrafted distortion aware + self-supervised pre-trained content features	HoloQA	SVR	**0.9144**	**0.7443**	**0.9489**	**0.2671**	**0.9201**	**0.7506**	**0.9523**	**0.2650**
	HoloQA+ 🔥	SVR	**0.9165**	**0.7507**	**0.9514**	**0.2598**	**0.9254**	**0.7641**	**0.9583**	**0.2457**

were randomly chosen from 80% of the content to form the training and validation sets, with the remaining 20% forming the test set. We also maintained content separation between the training and validation sets to prevent biases arising from content learning.

B. Objective Performance Comparisons of HoloQA FR-VQA Models

We benchmarked 15 widely-used 2D FR IQA/VQA algorithms and our proposed *HoloQA* models on the LIVE-Meta Rendered Human Avatar VQA database. The spectrum of compared algorithms extends from traditional FR-IQA/VQA models including PSNR (both RGB & Y channel versions), SSIM, DLM, VIF, VMAF (original and retrained versions), ST-GREED, SpEED-QA, ForVideoVDP, FUNQUE (and its enhanced versions), to more approaches using deep learning including LPIPS (AlexNet and VGG backbones), CONTRIQUE-FR, Re-IQA-FR, and CONVIQT-FR.

The algorithms PSNR, SSIM, DLM, VIF, and SpEED-QA operate without training and were thus directly applied to the 1000 test splits. For frame-based (FR-IQA) algorithms like PSNR, SSIM, DLM, VIF, CONTRIQUE-FR, Re-IQA-FR, and LPIPS, features or predicted scores were gathered on a per-frame basis, then averaged to yield video-level features or scores. To accomplish unbiased benchmarking, features from all algorithms necessitating training were gathered and mapped to the DMOS over 1000 training splits employing a Support Vector Regressor (SVR). The trained SVR model was then used to evaluate each algorithm's performance on the 1000 test splits.

The results in Table II summarize the objective performance of the FR IQA/VQA models on the LIVE-Meta Rendered Human Avatar database using body and face bounding boxes following evaluation strategy in [12]. We also evaluate *HoloQA* on full frames without computing bounding boxes; these results can be found in Section VA of the Supplementary material. Classic FR IQA frame-based models like PSNR, SSIM, DLM, and VIF yielded moderately inferior performance, while traditional FR VQA models like VMAF, ST-GREED, SpEED-QA, and FUNQUE obtained improvements over the frame-based FR IQA models, highlighting the importance of perceptual temporal modeling. The FovVideoVDP model with default calibration performed poorly on the LIVE-Meta Rendered Human Avatar dataset but showed improvement with retraining, as discussed later, in Section VIB. We present results for both the original and retrained versions of VMAF, observing a slight enhancement with retraining. FUNQUE and its refined variants, initially designed to enhance the time complexity of VMAF, exhibit anticipated slight performance improvements over VMAF. However, the upgraded versions of FUNQUE, namely Y-FUNQUE+ and 3C-FUNQUE+ [21] yielded similar performance as the original model on the LIVE-Meta Rendered Human Avatar dataset. This can be attributed to the features used in models targeting scaling and compression distortions, which differ from many of the distortions arising in our case. For the family of models using entropy-based statistics, SpEED-QA yielded sub-optimal performance, likely since it is non-trainable. ST-GREED which uses powerful temporal model of distortion perception, performed quite well. The frame-based learning methods LPIPS, CONTRIQUE-FR, and ReIQA-FR yielded superior performance, sometimes surpassing the performance of the FR-VQA models, even without the advantage of temporal processing. CONVIQT-FR obtains strong but slightly lower

TABLE III
EFFECT OF ADDING DENSE FRAME TRACKING ON THE MEDIAN PERFORMANCE OF HOLOQA VARIANTS ON THE LIVE-META RENDERED HUMAN AVATAR VQA DATABASE (1000 TRAIN/TEST SPLITS). ENHANCED VERSIONS OF *HoloQA* ARE INDICATED WITH 🔥. BEST-PERFORMING CONFIGURATION IS BOLDFACED.

Full Reference Algorithm		Regressor	Body Bounding Box				Face Bounding Box			
			SRCC (↑)	KRCC (↑)	PLCC (↑)	RMSE (↓)	SRCC (↑)	KRCC (↑)	PLCC (↑)	RMSE (↓)
Handcrafted distortion-aware features + self-supervised pre-trained deep content features	HoloQA	SVR	0.9144	0.7443	0.9489	0.2671	0.9201	0.7506	0.9523	0.2650
	HoloQA with Frame Tracking 🔥	SVR	0.9163	0.7497	0.9531	0.2573	0.9291	0.7709	0.9593	0.2497
	HoloQA+ 🔥	SVR	0.9165	0.7507	0.9514	0.2598	0.9254	0.7641	0.9583	0.2457
	HoloQA+ with Frame Tracking 🔥🔥	SVR	**0.9201**	**0.7517**	**0.9556**	**0.2512**	**0.9350**	**0.7783**	**0.9618**	**0.2432**

performance than CONTRIQUE. Finally, *HoloQA/HoloQA+* demonstrates superior performance compared to all other models by leveraging both distortion and content-aware modeling for overall video quality prediction. The effect incorporating dense frame-tracking incorporated in *HoloQA/HoloQA+* variants are provided in Table III. The benefits of using frame-wise human body/face tracking and human body/face-specific content-aware fine-tuning are evident from the variants of *HoloQA* annotated with 🔥 in Tables II,III. We provide further insights into the contributions of the different components of *HoloQA* in Section VI.

VI. ABLATION STUDY

A. Effects of Bounding Boxes & Frame Tracking on Distortion-Aware Module

We investigated the effects of frame tracking and bounding box type on the HoloQA by testing various combinations: frame tracking on/off and face versus body bounding boxes. The results in Table IV show that the configuration using dense frame tracking delivered the best performance when either face or body bounding boxes were used.

TABLE IV
EFFECTS OF BOUNDING BOXES AND FRAME TRACKING ON THE DISTORTION-AWARE MODULE OF HOLOQA. MEDIAN SRCC/PLCC OVER 1000 TRAIN/TEST SPLITS. BEST PER BOUNDING BOX IN BOLD.

Bounding Box	Frame Tracking: On		Frame Tracking: Off	
	SRCC	PLCC	SRCC	PLCC
Body	**0.9113**	**0.9473**	0.9079	0.9428
Face	**0.9253**	**0.9587**	0.9188	0.9518

B. Effects of Distortion Aware Feature Components

Next, we studied the impacts of the three categories of distortion-aware features employed in the Distortion Aware module of HoloQA. We evaluate them individually, in pairs, and all three feature sets together using the best-performing configuration found in Section VI-A. Table V summarizes the results. An interesting phenomenon emerges: although the performance of the entropy differencing feature map is poor on its own, it noticeably enhances performance when combined with one or both of the other distortion-aware features. Furthermore, we conducted two-sample one-sided t-tests using the 1000 SROCC and PLCC values from the configuration using all three categories of features and comparing them with the other configurations in Table V. The results show that the differences were statistically significant, demonstrating that combining all three features improves the model's performance with statistical significance. It is also worth noting that, employing only the contrast masking (CM) features with a trained regressor on the LIVE-Meta Rendered Human dataset may be regarded as a simplified proxy to optimize the pooling and normalization processes in the model [22] to obtain dataset-specific fine-tuning. As expected, this adaptation yields improved performance as compared to the original calibrated [22] model across a larger collection of datasets. In Section VB of the Supplementary Material, we also discuss the performance of each distortion-aware feature combination across distortions.

C. Effects of Bounding Box and Frame Tracking on Content-Aware Module of Holo-QA

Similar to Section VI-A, where we conducted an ablation study on the design parameters of the Distortion Aware module of HoloQA, we also performed an ablation study on the factors affecting the Content-Aware Module. Specifically, we investigate the effects of enabling or disabling frame tracking, the impact of the bounding boxes, and whether the self-supervised pretrained model is trained on the generic ImageNet dataset or fine-tuned on a content-specific dataset of human bodies or faces. The results in Table VI indicate that the configuration utilizing dense frame tracking and models fine-tuned on content-specific datasets performed the best for both face and body bounding boxes.

D. Distortion Aware v/s Content Aware

In this section, we analyze the performance of the Hologram-Distortion aware and Content Aware Modules of *HoloQA* individually and in combination. We employed dense frame tracking for the body and face bounding boxes. The results in Table VII demonstrate the performance improvements achieved through the *Mixture-of-Experts* approach, which leverages distortion-aware and content-aware features together.

VII. CONCLUSION & FUTURE WORK

This paper introduces HoloQA, a Full Reference Video Quality Assessment model that leverages insights from recent developments in visual neuroscience, information theory, and self-supervised deep learning. It is able to accurately predict the perceptual quality of rendered digital human avatar videos

TABLE V
DISTORTION-AWARE MODULE COMPONENTS OF HOLOQA. MEDIAN SRCC/PLCC OVER 1000 TRAIN/TEST SPLITS. BEST PERFORMANCE IS BOLDFACED.

Feature Set	Bounding Box: Body		Bounding Box: Face	
	SRCC	PLCC	SRCC	PLCC
Structural Similarity (SS)	0.9017	0.9436	0.9121	0.9482
Entropic Differencing (ED)	0.6145	0.6560	0.6988	0.7582
Contrast Masking (CM)	0.8512	0.8716	0.9072	0.9467
SS + ED	0.9052	0.9465	0.9219	0.9485
SS + CM	0.9089	0.9458	0.9151	0.9522
ED + CM	0.9020	0.9321	0.9170	0.9501
SS + ED + CM	**0.9113**	**0.9473**	**0.9253**	**0.9587**

TABLE VI
EFFECTS OF BOUNDING BOX AND FRAME TRACKING ON THE CONTENT-AWARE MODULE OF HOLOQA. MEDIAN SRCC/PLCC OVER 1000 TRAIN/TEST SPLITS. BEST PERFORMANCES FOR EACH BOUNDING BOX ARE BOLDFACED.

Configuration	Frame Tracking: On		Frame Tracking: Off	
	SRCC	PLCC	SRCC	PLCC
HoloQA–Content Aware: Body	0.8963	0.9350	0.8850	0.9177
HoloQA+–Content Aware: Body	**0.9039**	**0.9427**	0.8932	0.9280
HoloQA–Content Aware: Face	0.8799	0.9053	0.8530	0.8760
HoloQA+–Content Aware: Face	**0.8896**	**0.9107**	0.8594	0.8835

such as those that occur in Virtual Reality (VR) and Augmented Reality (AR) environments. We compare the superior performances of many leading generic and dedicated FR VQA models on the LIVE-Meta Rendered Human Avatar VQA database and find all versions of *HoloQA* deliver superior performance. In the Supplementary material, we also studied the performance of *HoloQA* on other digital human avatar databases and on a synthetically generated Cloud Gaming dataset.

TABLE VII
ABLATION ON DISTORTION-AWARE AND CONTENT-AWARE MODULES OF HOLOQA. MEDIAN SRCC/PLCC OVER 1000 TRAIN/TEST SPLITS. BEST PERFORMANCE PER COLUMN IS BOLDFACED. RESULTS ARE WITH FRAME TRACKING ENABLED.

Feature Set	Bounding Box: Body		Bounding Box: Face	
	SRCC	PLCC	SRCC	PLCC
HoloQA (Distortion Aware Only)	0.9113	0.9473	0.9253	0.9587
HoloQA (Content Aware Only)	0.8963	0.9350	0.8799	0.9053
HoloQA+ (Content Aware Only)	0.9039	0.9427	0.8896	0.9107
HoloQA (Content + Distortion Aware)	0.9163	0.9531	0.9291	0.9593
HoloQA+ (Content + Distortion Aware)	**0.9201**	**0.9556**	**0.9350**	**0.9618**

REFERENCES

[1] "Virtual Reality (VR) Market Size, Share & Trends Analysis Report By Technology (Semi & Fully Immersive, Non-immersive), By Device (HMD, GTD), By Component (Hardware, Software), By Application, By Region, And Segment Forecasts, 2023 - 2030," https://www.grandviewresearch.com/industry-analysis/virtual-reality-vr-market, 2022, [Online; accessed 15-November-2023].

[2] K. Li, X. Peng, J. Song, B. Hong, and J. Wang, "The application of augmented reality (ar) in remote work and education," 2024. [Online]. Available: https://zenodo.org/doi/10.5281/zenodo.10965828

[3] Z. Wang, L.-P. Yuan, L. Wang, B. Jiang, and W. Zeng, "Virtuwander: Enhancing multi-modal interaction for virtual tour guidance through large language models," vol. 4, p. 1–20, May 2024. [Online]. Available: http://dx.doi.org/10.1145/3613904.3642235

[4] "Meta Horizon Workrooms," https://forwork.meta.com/horizon-workrooms/, 2023, [Online; accessed 15-November-2023].

[5] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[6] Z. Wang, E. Simoncelli, and A. Bovik, "Multiscale structural similarity for image quality assessment," in *Asilomar Conference on Signals, Systems & Computers*, vol. 2, 2003, pp. 1398–1402.

[7] Z. Li, A. Aaron, I. Katsavounidis, A. Moorthy, and M. Manohara, "Toward a practical perceptual video quality metric," *The Netflix Tech Blog*, vol. 6, no. 2, 2016.

[8] I. Katsavounidis, "Dynamic optimizer — a perceptual video encoding optimization framework," *The Netflix Tech Blog*, vol. 6, no. 2, 2018.

[9] Z. Jiang, K. Venkateshan, G. Nam, M. Chen, R. Bachy, J.-C. Bazin, and A. Chapiro, "Facemap: Distortion-driven perceptual facial saliency maps," pp. 1–11, 2024.

[10] M. Ashraf, A. Chapiro, and R. K. Mantiuk, "Resolution limit of the eye: how many pixels can we see?" 2024.

[11] A. Saha, S. K. Pentapati, Z. Shang, R. Pahwa, B. Chen, H. E. Gedik, S. Mishra, and A. C. Bovik, "Perceptual Video Quality Assessment: The Journey Continues!" *Frontiers in Signal Processing*, vol. 3, 2023. [Online]. Available: https://www.frontiersin.org/articles/10.3389/frsip.2023.1193523

[12] Y.-C. Chen, A. Saha, A. Chapiro, C. Häne, J.-C. Bazin, B. Qiu, S. Zanetti, I. Katsavounidis, and A. C. Bovik, "Subjective and objective quality assessment of rendered human avatar videos in virtual reality," *IEEE Transactions on Image Processing*, vol. 33, p. 5740–5754, 2024. [Online]. Available: http://dx.doi.org/10.1109/tip.2024.3468881

[13] Z. Wang and A. C. Bovik, "Mean squared error: Love it or leave it? a new look at signal fidelity measures," *IEEE Signal Processing Magazine*, vol. 26, no. 1, pp. 98–117, 2009.

[14] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "Fsim: A feature similarity index for image quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.

[15] H. Sheikh and A. Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, 2006.

[16] R. Soundararajan and A. Bovik, "Video quality assessment by reduced reference spatio-temporal entropic differencing," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, pp. 684–694, 2013.

[17] C. G. Bampis, P. Gupta, R. Soundararajan, and A. C. Bovik, "Speed-qa: Spatial efficient entropic differencing for image and video quality," *IEEE Signal Processing Letters*, vol. 24, no. 9, pp. 1333–1337, 2017.

[18] P. C. Madhusudana, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik, "St-greed: Space-time generalized entropic differences for frame rate dependent video quality prediction," *IEEE Transactions on Image Processing*, vol. 30, pp. 7446–7457, 2021.

[19] S. Li, F. Zhang, L. Ma, and K. N. Ngan, "Image quality assessment by separately evaluating detail losses and additive impairments," *IEEE Transactions on Multimedia*, vol. 13, no. 5, pp. 935–949, 2011.

[20] A. K. Venkataramanan, C. Stejerean, and A. C. Bovik, "FUNQUE: Fusion of unified quality evaluators," 2022. [Online]. Available: https://arxiv.org/abs/2202.11241

[21] A. K. Venkataramanan, C. Stejerean, I. Katsavounidis, and A. C. Bovik, "One transform to compute them all: Efficient fusion-based full-reference video quality assessment," *IEEE Transactions on Image Processing*, vol. 33, pp. 509–524, 2024.

[22] R. K. Mantiuk, G. Denes, A. Chapiro, A. Kaplanyan, G. Rufo, R. Bachy, T. Lian, and A. Patney, "Fovvideovdp: A visible difference predictor for wide field-of-view video," *ACM Trans. Graph.*, vol. 40, no. 4, jul 2021.

[23] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Los Alamitos, CA, USA, 2018, pp. 586–595.

[24] W. Kim, J. Kim, S. Ahn, J. Kim, and S. Lee, "Deep video quality assessor: From spatio-temporal visual sensitivity to a convolutional neural aggregation network," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 219–234.

[25] M. Xu, J. Chen, H. Wang, S. Liu, G. Li, and Z. Bai, "C3dvqa: Full-reference video quality assessment with 3d convolutional neural network," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 4447–4451.

[26] Y. Li, L. Feng, J. Xu, T. Zhang, Y. Liao, and J. Li, "Full-reference and no-reference quality assessment for compressed user-generated content videos," in *IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, 2021, pp. 1–6.

[27] P. C. Madhusudana, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C.

Bovik, "Image quality assessment using contrastive learning," *IEEE Transactions on Image Processing*, vol. 31, pp. 4149–4161, 2022.
[28] P. Madhusudana, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik, "Conviqt: Contrastive video quality estimator," *IEEE Transactions on Image Processing*, vol. 32, pp. 5138–5152, 2023.
[29] A. Saha, S. Mishra, and A. C. Bovik, "Re-IQA: Unsupervised learning for image quality assessment in the wild," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2023, pp. 5846–5855.
[30] W. Wen, Y. Wu, Y. Sheng, N. Birkbeck, B. Adsumilli, and Y. Wang, "Cp-llm: Context and pixel aware large language model for video quality assessment," *arXiv*, 2025. [Online]. Available: https://arxiv.org/abs/2505.16025
[31] H. Wu, Z. Zhang, W. Zhang, C. Chen, L. Liao, C. Li, Y. Gao, A. Wang, E. Zhang, W. Sun, Q. Yan, X. Min, G. Zhai, and W. Lin, "Q-align: Teaching lmms for visual scoring via discrete text-defined levels," *arXiv*, 2023.
[32] H. Wu, Z. Zhang, E. Zhang, C. Chen, L. Liao, A. Wang, K. Xu, C. Li, J. Hou, G. Zhai, G. Xue, W. Sun, Q. Yan, and W. Lin, "Q-instruct: Improving low-level visual abilities for multi-modality foundation models," *arXiv*, 2023.
[33] L. Dong, Y. Fang, W. Lin, and H. S. Seah, "Perceptual quality assessment for 3d triangle mesh based on curvature," *IEEE Transactions on Multimedia*, vol. 17, no. 12, pp. 2174–2184, 2015.
[34] K. Wang, F. Torkhani, and A. Montanvert, "A fast roughness-based approach to the assessment of 3D mesh visual quality," *Computers Graphics*, vol. 36, no. 7, pp. 808–818, 2012.
[35] G. Lavoué, "A multiscale metric for 3D mesh visual quality assessment," *Comput. Graph. Forum*, vol. 30, pp. 1427–1437, 08 2011.
[36] S. Yang, C.-H. Lee, and C.-C. J. Kuo, "Optimized mesh and texture multiplexing for progressive textured model transmission," in *ACM International Conference on Multimedia*, 2004, p. 676–683.
[37] F. Caillaud, V. Vidal, F. Dupont, and G. Lavoué, "Progressive compression of arbitrary textured meshes," in *Computer Graphics Forum*, vol. 35, 10 2016.
[38] A. Nouri, C. Charrier, and O. Lézoray, "Full-reference saliency-based 3d mesh quality assessment index," in *2016 IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 1007–1011.
[39] Y. Nehmé, J. Delanoy, F. Dupont, J.-P. Farrugia, P. Le Callet, and G. Lavoué, "Textured mesh quality assessment: Large-scale dataset and deep learning-based quality metric," *ACM Transactions on Graphics*, vol. 42, no. 3, p. 1–20, Jun. 2023.
[40] I. Abouelaziz, A. Chetouani, M. El Hassouni, and H. Cherifi, "A blind mesh visual quality assessment method based on convolutional neural network," *Electronic Imaging*, vol. 2018, pp. 423–1, 01 2018.
[41] Z. Zhang, Y. Zhou, W. Sun, X. Min, Y. Wu, and G. Zhai, "Perceptual quality assessment for digital human heads," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Jun. 2023.
[42] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, oct 2021, pp. 9992–10 002.
[43] Z. Zhang, W. Sun, Y. Zhou, H. Wu, C. Li, X. Min, X. Liu, G. Zhai, and W. Lin, "Advancing zero-shot digital human quality assessment through text-prompted evaluation," *CoRR*, vol. abs/2307.02808, 2023.
[44] Z. Zhang, Y. Zhou, W. Sun, W. Lu, X. Min, Y. Wang, and G. Zhai, "Ddh-qa: A dynamic digital humans quality assessment database," in *2023 IEEE International Conference on Multimedia and Expo (ICME)*, 2023, pp. 2519–2524.
[45] K. Wolski, L. Trutoiu, Z. Dong, Z. Shen, K. MacKenzie, and A. Chapiro, "Geo-metric: A perceptual dataset of distortions on faces," *ACM Transactions on Graphics (TOG)*, vol. 41, 2022.
[46] "Meta Quest Pro," https://www.meta.com/quest/quest-pro/, 2023, [Online; accessed 15-November-2023].
[47] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.
[48] "Perceptual Display Calibration," https://www.cl.cam.ac.uk/~rkm38/pdfs/mantiuk2016perceptual_display.pdf, 2016, [Online; accessed 15-November-2023].
[49] M. Livingstone and D. Hubel, "Segregation of form, color, movement, and depth: anatomy, physiology, and perception," *Science*, vol. 240, no. 4853, pp. 740–749, 1988.
[50] D. H. Hubel and T. N. Wiesel, "Receptive fields and functional architecture in two nonstriate visual areas (18 and 19) of the cat," *Journal of neurophysiology*, vol. 28, pp. 229–289, 1965.
[51] D. C. Van Essen and J. L. Gallant, "Neural mechanisms of form and motion processing in the primate visual system," *Neuron*, vol. 13, no. 1, pp. 1–10, 1994.
[52] A. M. Derrington and P. Lennie, "Spatial and temporal contrast sensitivities of neurones in lateral geniculate nucleus of macaque," *The Journal of physiology*, vol. 357, pp. 219–240, 1984.
[53] P. Burt and E. Adelson, "The laplacian pyramid as a compact image code," *IEEE Transactions on Communications*, vol. 31, no. 4, pp. 532–540, 1983.
[54] T. Tariq, N. Matsuda, E. Penner, J. Jia, D. Lanman, A. Ninan, and A. Chapiro, "Perceptually adaptive real-time tone mapping," in *SIGGRAPH Asia 2023 Conference Papers*, 2023, pp. 1–10.
[55] F. A. Kingdom and P. Whittle, "Contrast discrimination at high contrasts reveals the influence of local light adaptation on contrast processing," *Vision Research*, vol. 36, no. 6, pp. 817–829, 1996.
[56] M. P. Sampat, Z. Wang, S. Gupta, A. C. Bovik, and M. K. Markey, "Complex wavelet structural similarity: A new image similarity index," *IEEE Transactions on Image Processing*, vol. 18, no. 11, pp. 2385–2401, 2009.
[57] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Transactions on Image Processing*, vol. 19, no. 2, pp. 335–350, 2010.
[58] Z. Tu, X. Yu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "RAPIQUE: Rapid and accurate video quality prediction of user generated content," *IEEE Open Journal of Signal Processing*, vol. 2, p. 425–440, 2021.
[59] Y.-C. Chen, A. Saha, C. Davis, B. Qiu, X. Wang, R. Gowda, I. Katsavounidis, and A. C. Bovik, "GAMIVAL: Video quality prediction on mobile cloud gaming content," *IEEE Signal Processing Letters*, vol. 30, p. 324–328, 2023.
[60] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 9620–9629.
[61] J. Fu, S. Li, Y. Jiang, K.-Y. Lin, C. Qian, C.-C. Loy, W. Wu, and Z. Liu, "Stylegan-human: A data-centric odyssey of human generation," *arXiv preprint*, vol. arXiv:2204.11823, 2022.
[62] "VGGFace2-HQ," https://github.com/NNNNAI/VGGFace2-HQ, [Online; accessed 2-June-2024].
[63] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. H. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. D. Guo, M. G. Azar, B. Piot, K. Kavukcuoglu, R. Munos, and M. Valko, "Bootstrap your own latent: A new approach to self-supervised learning," *CoRR*, vol. abs/2006.07733, 2020.
[64] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *CoRR*, vol. abs/2003.04297, 2020.
[65] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *CoRR*, vol. abs/1807.03748, 2018.
[66] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, p. 1–1, 2020. [Online]. Available: http://dx.doi.org/10.1109/TPAMI.2020.3045810
[67] E. Alexiou and T. Ebrahimi, "Exploiting user interactivity in quality assessment of point cloud imaging," in *Quality of Multimedia Experience (QoMEX)*, 2019, pp. 1–6.
[68] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "Yolov7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 7464–7475.
[69] "YOLO by Ultralytics," https://github.com/ultralytics/ultralytics, [Online; accessed 7-March-2024].
[70] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, "Study of subjective and objective quality assessment of video," *IEEE Transactions on Image Processing*, vol. 19, no. 6, pp. 1427–1441, 2010.