# No-Reference Rendered Video Quality Assessment: Dataset and Metrics

Sipeng Yang, Jiayu Ji, Qingchuan Zhu, Zhiyao Yang, Xiaogang Jin, *Member, IEEE*

***Abstract*—Quality assessment of videos is crucial for many computer graphics applications, including video games, virtual reality, and augmented reality, where visual performance has a significant impact on user experience. When test videos cannot be perfectly aligned with references or when references are unavailable, the significance of no-reference video quality assessment (NR-VQA) methods is undeniable. However, existing NR-VQA datasets and metrics are primarily focused on camera-captured videos; applying them directly to rendered videos would result in biased predictions, as rendered videos are more prone to temporal artifacts. To address this, we present a large rendering-oriented video dataset with subjective quality annotations, as well as a designed NR-VQA metric specific to rendered videos. The proposed dataset includes a wide range of 3D scenes and rendering settings, with quality scores annotated for various display types to better reflect real-world application scenarios. Building on this dataset, we calibrate our NR-VQA metric to assess rendered video quality by looking at both image quality and temporal stability. We compare our metric to existing NR-VQA metrics, demonstrating its superior performance on rendered videos. Finally, we demonstrate that our metric can be used to benchmark supersampling methods and assess frame generation strategies in real-time rendering. The dataset and calibrated metric are publicly available at the project homepage `https://mob-fgsr.github.io/ReVQ/`.**

***Index Terms*—Video quality assessment, rendered video evaluation, rendering artifacts.**

## I. Introduction

VIDEO quality metrics are essential for optimizing rendering pipelines to ensure high-fidelity outcomes in rendered content [1]. Well-known metrics, such as structural similarity (SSIM) [2] and peak signal-to-noise ratio (PSNR), along with human visual system (HVS)-based methods [1], [3], require perfectly aligned and high-quality reference images to evaluate the similarity between test images and references. However, in commercial graphics rendering applications, misalignment between test and reference images frequently occurs due to the difficulty in consistently replicating exact camera positions and dynamic object poses across different rendering cycles, which reduces the effectiveness of full-reference metrics. These issues highlight the critical need for reliable no-reference video quality assessment (NR-VQA) metrics.

NR-VQA methods aim to evaluate the perceptual quality of test videos without relying on references. Extensive research has been conducted in this field, with datasets such as KoNViD-1k [4] and LIVE-VQC [5], as well as developed NR-VQA metrics focusing on temporal aggregation [6], [7], and temporal pooling [8], [9]. These efforts primarily target the quality assessment of camera-captured videos, which are usually degraded by *blurring*, *noise*, and *overexposure* issues [10]. However, the main factors affecting the perceptual quality of rendered videos differ from those impacting camera-captured ones. Rendered videos, limited by per-pixel sampling rates, are particularly susceptible to temporal artifacts like *flickering* and *moving jaggies*. These artifacts are especially pronounced and disruptive due to the human visual system's acute sensitivity to temporal variations [11]. As a result, there is an urgent need to create a new dataset and a specifically designed metric for NR-VQA of rendered videos.

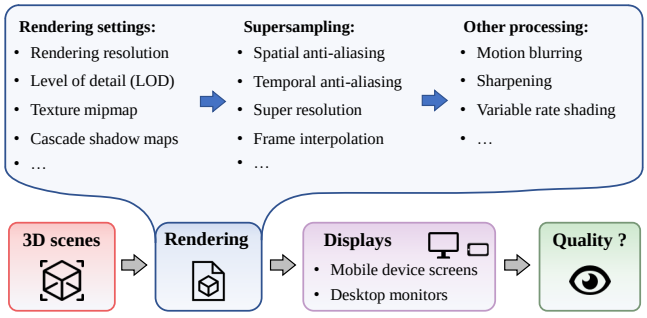



Fig. 1. Process for assessing the quality of rendered videos. Rendered videos undergo a complex transition from 3D scenes to on-screen displays. During this transition, 3D scene quality, rendering configurations, and display types all have a significant impact on perceived quality.

Developing NR-VQA methods for rendered videos first requires addressing the scarcity of suitable datasets. The perceived quality of rendered videos is influenced by multiple factors, including 3D scene resources, rendering pipelines, and display devices, as shown in Fig. 1. Ensuring these aspects are represented in the dataset is crucial to effectively reflect real-world conditions. Besides the dataset, designing effective NR-VQA metrics for rendered videos poses additional challenges. For rendered content, recent convolutional neural network (CNN)-based models [12] have leveraged extra G-buffers to detect *static artifacts*, such as aliasing and Moiré effects. However, detecting *temporal artifacts* or assessing temporal stability in dynamic video scenes, especially in the absence of reference images, remains a significant unresolved challenge. As illustrated in Fig. 7, failing to consider the temporal stability of videos can result in significant biases in NR-VQA of rendered videos.

To address the challenges, we introduce a large rendering-

Sipeng Yang, Jiayu Ji, Qingchuan Zhu, and Xiaogang Jin are with the State Key Lab of CAD&CG, Zhejiang University, Hangzhou, P. R. China.
Zhiyao Yang is with OPPO Nanjing Research Center, Nanjing, P. R. China.
Xiaogang Jin is the corresponding author. E-mail: jin@cad.zju.edu.cn
Manuscript received MM DD, 2024; revised MM DD, 2024.

oriented dataset accompanied by a new NR-VQA metric tailored for rendered videos. The dataset, named *ReVQ-2k*, features a diverse array of 3D scenes and rendering configurations, with perceptual quality scores annotated on both smartphone screens and desktop displays. To enhance the evaluation on video temporal stability, we expand our collection of subjective annotations to include not only overall quality scores but also temporal stability scores, providing additional supervision for metric calibration. Simultaneously, we propose an NR-VQA metric that evaluates rendered video quality from two perspectives. First, we focus on image quality factors such as clarity and static artifacts, aligning with existing NR-VQA methods [9], [13]. Given the extensive analysis in existing literature, we directly adopt a state-of-the-art (SOTA) practice, FAST-VQA [13], for the assessment of the image quality aspect. Second, for the evaluation of temporal artifacts, we propose utilizing motion estimation to counteract object movement across frames, and then using a multi-timescale image differencing module to assess the temporal stability of videos. This module can be calibrated using the annotated temporal stability scores from ReVQ-2k to achieve better precision. Finally, our metric integrates these two aspects to provide a comprehensive evaluation of rendered video quality, offering more accurate predictions than existing metrics.

We demonstrate the utility of our calibrated NR-VQA metric through practical applications, including comparing the video quality of closed-source supersampling methods and assessing the perceived quality of frame generation strategies. Our metric provides stable quantitative assessments, offering substantial advantages over manual annotations for many relevant real-world applications.

The main contributions of the paper are as follows:

- We introduce a large rendering-oriented dataset, *ReVQ-2k*, comprising 2,000 rendered videos that feature a wide range of scenes and rendering settings, along with 57,450 subjective annotations for overall quality and temporal stability on various displays.
- We develop a novel NR-VQA metric that considers both image quality and temporal stability of rendered videos.
- The utility of the proposed metric is demonstrated in real-world tasks, including the evaluation of closed-source supersampling methods and frame generation strategies.

## II. Related Works

### A. NR-VQA Datasets

High-quality datasets are essential for the calibration and evaluation of NR-VQA metrics. Early datasets [14], [15] in this field primarily focus on videos affected by specific distortions from compression and transmission processes. These datasets are limited to a narrow range of distortions, which restricts their utility in real-world scenarios [16]. In contrast, subsequent studies like CVD2014 [17], LIVE-Qualcomm [18], and KoNViD-1k [4] have collected a wide variety of videos and provided manual quality ratings, significantly enhancing the diversity and practical relevance of these resources. These datasets have become instrumental in advancing NR-VQA research, particularly for assessing real-world video quality. However, rendered videos, which are increasingly prevalent in various applications, are scarcely represented within these datasets.

Despite a recent work, LIVE-YT-Gaming [19], introduces a dataset for video streams VQA of gaming content, it primarily evaluates the impact of video compression, without considering rendering settings and display impacts as illustrated in Fig. 1. Our work aims to fill this gap by collecting data with a broad range of 3D scenes and rendering settings, along with providing quality ratings for videos displayed on various types of screens. This initiative significantly enhances the capability of NR-VQA metrics to assess the quality of rendered videos, providing crucial resources for quality evaluation in this field.

### B. NR-VQA metrics

**Classical NR-VQA Methods.** Over the past decades, extensive efforts in NR-VQA have been conducted from diverse perspectives. A prominent strategy involves quantifying artifacts in test videos. Early studies thoroughly examine the impact of image artifacts such as blocking [20], noise [21], and blurring [22] on perceived video quality. Additionally, alternative approaches assess video quality using visual indicators like local contrast, brightness, and colorfulness [23], [24]. These NR-VQA approaches leverage hand-crafted features, which enhance the model interpretability. However, they often yield biased evaluations when confronted with diverse video content and distortion types.

**Deep Learning-Based NR-VQA Methods.** To improve evaluation performance across various video categories, recent NR-VQA methods have shifted towards leveraging large video datasets with manually annotated quality labels to train deep learning models for automatic prediction of perceived video quality. Due to the superior feature extraction capabilities of deep neural networks (DNNs), models based on these networks generally outperform those relying on hand-crafted features in terms of accuracy. Among the various techniques employed, the gated recurrent unit (GRU)-based module, capable of learning long-term dependencies in sequential data, is prevalent in NR-VQA algorithms to model the temporal features of videos [10], [25]. Other temporal modeling methods, including pyramid temporal aggregation [6], motion features statistic [26], 3D CNNs [27], and SlowFast networks [28], have also proven effective in NR-VQA for temporal information modeling.

Extensive experiments have demonstrated the effectiveness of DNNs-based NR-VQA metrics on real-world datasets [9]. However, these methods are generally not designed to account for rendering-specific artifacts such as moving jaggies and flickering [29], [30], which are prevalent in rendered videos. These limitations motivate us to develop a new NR-VQA metric that is specifically designed for predicting rendered video quality.

### C. Full-Reference Metrics for Rendered Videos

In addition to no-reference solutions, full-reference metrics have been specifically designed and widely used to evaluate

TABLE I
SUMMARY OF EXISTING NR-VQA DATASETS AND THE PROPOSED REVQ-2K DATASET.

| Dataset | Scale | Duration | Resolution | Frame Rate | Content Origin |
|---|---|---|---|---|---|
| CVD2014 [17] | 234 | 10-25 sec | 480p, 720p | 10-32 | Captured with real cameras, no post-processing distortions. |
| LIVE-Qualcomm [18] | 208 | 15 sec | 1080p | 30 | Captured with mobile devices. |
| KoNViD-1k [4] | 1,200 | 8 sec | 540p | 24, 25, 30 | Camera-captured in-the-wild videos. |
| LIVE-VQC [5] | 585 | 10 sec | 240p-1080p | 20-30 | Captured using 101 different camera devices. |
| LIVE-YT-Gaming [19] | 600 | 8-9 sec | 360p-1080p | 30, 60 | Compressed streaming game videos. |
| **ReVQ-2k (ours)** | 2,000 | 8 sec | 720p, 1080p, 2k | 60 | Rendered videos using various rendering settings, evaluated on different types of displays. |

rendered video quality. These methods can be broadly categorized into two types: similarity measures and artifact detection-based approaches. Similarity measures, such as SSIM [2], PSNR, and root mean square error (RMSE), operate on the assumption that greater similarity to a reference image indicates higher content quality, whereas artifact detection-based methods concentrate on the impact of specific artifacts on rendered video quality. Due to the scarcity of training data with manual annotations, methods like the contrast sensitivity function (CSF) [31] are frequently employed for visual distortion detection. For instance, Aydin *et al.* [32] use a 3D CSF and psychometric function metrics to assess distortion visibility in computer-generated videos. Another study [33] proposes a calibrated human visual system model for predicting distortion maps in high dynamic range images. Mantiuk *et al.* [3] have developed a per-pixel visual difference predictor to compare reference and distorted video sequences. More recently, deep learning networks have been employed to detect artifacts in rendered content by training on image datasets with localized distortion maps for accurate visible distortion prediction [12], [34]. However, these methods depend on ground-truth references or rendered G-buffers and fail to assess the temporal stability of videos [35], leaving a gap in research specifically for NR-VQA of rendered videos.

## III. REVQ-2K DATASET

We begin by introducing the proposed rendered video dataset and the subjective quality study conducted on it. Our dataset, known as rendered video quality-2k (ReVQ-2k), includes 2,000 rendered video clips and 57,450 subjective quality annotations. Tab. I provides a detailed comparison of existing datasets.

### A. Video Collection

Our data collection strategy is guided by the analysis presented in Fig. 1, which outlines the origins of rendered videos. This strategy incorporates a wide range of 3D scenes and objects, resulting in a rich diversity of rendering styles and content. Moreover, our rendering pipeline utilizes a comprehensive suite of rendering configurations, supplemented by various supersampling and post-processing techniques, to accurately replicate real-world scenarios.

**3D Scenes.** The videos in our ReVQ-2k dataset are created using Unreal Engine 4 (UE4) [36] and Unreal Engine 5 (UE5) [37]. We select 15 diverse 3D scenes to encompass a broad range of visual environments. These environments include urban street scenes, interior settings, outdoor landscapes, and scenes rendered in a cartoon style. To further enhance diversity, we capture scenes at various times of the day, such as night and noon, as well as under different weather conditions, such as dusk and snow. Fig. 2 showcases examples of the scenes included in our dataset.

**Rendering Settings.** Unreal Engine allows for extensive rendering pipeline customization. When creating the ReVQ-2k dataset, we choose settings such as view distances, anti-aliasing methods, post-processing effects, shadows, textures, effect quality, and resolution scaling. These videos are generated at various scalability levels, with each using a random combination of adjustable settings. We also use popular super-sampling techniques for video generation, such as FidelityFX super resolution (FSR) [38], deep learning super sampling (DLSS) [39], and temporal anti-aliasing upscaling (TAAU) [40]. The project homepage includes detailed information about the rendering settings.

With the rendering settings established, we can now begin collecting rendered videos. We use three resolution settings on the selected 3D scenes: 720p, 1080p, and 2K. For each resolution setting, approximately 700 video clips are gathered. The 720p videos are optimized for smartphone screens, whereas the 1080p and 2K videos are designed for desktop monitors, reflecting typical real-world usage scenarios.

**Data Analysis.** To validate the diversity of the ReVQ-2k dataset, we analyze its low-level quantitative attributes [41] and compare them to those of established datasets. Attributes such as contrast, colorfulness, temporal information (TI), and brightness provide a comprehensive comparison of diversity among these datasets. Fig. 3 illustrates the distributions of these attributes across various datasets, including ReVQ-2k, CVD2014 [17], KoNViD-1k [4], LIVE-Qualcomm [18], LIVE-VQC [5], and LIVE-YT-Gaming [19]. Our analysis shows that ReVQ-2k features a wide range of contrast and colorfulness, aligning with existing datasets. It also exhibits high levels of TI, indicating more frequent and larger camera movements in ReVQ-2k videos. Although its brightness is slightly lower compared to other datasets, the difference is marginal. Given the rich and varied content of ReVQ-2k, we believe that it is well-suited for calibrating and evaluating VQA metrics, akin to existing datasets.

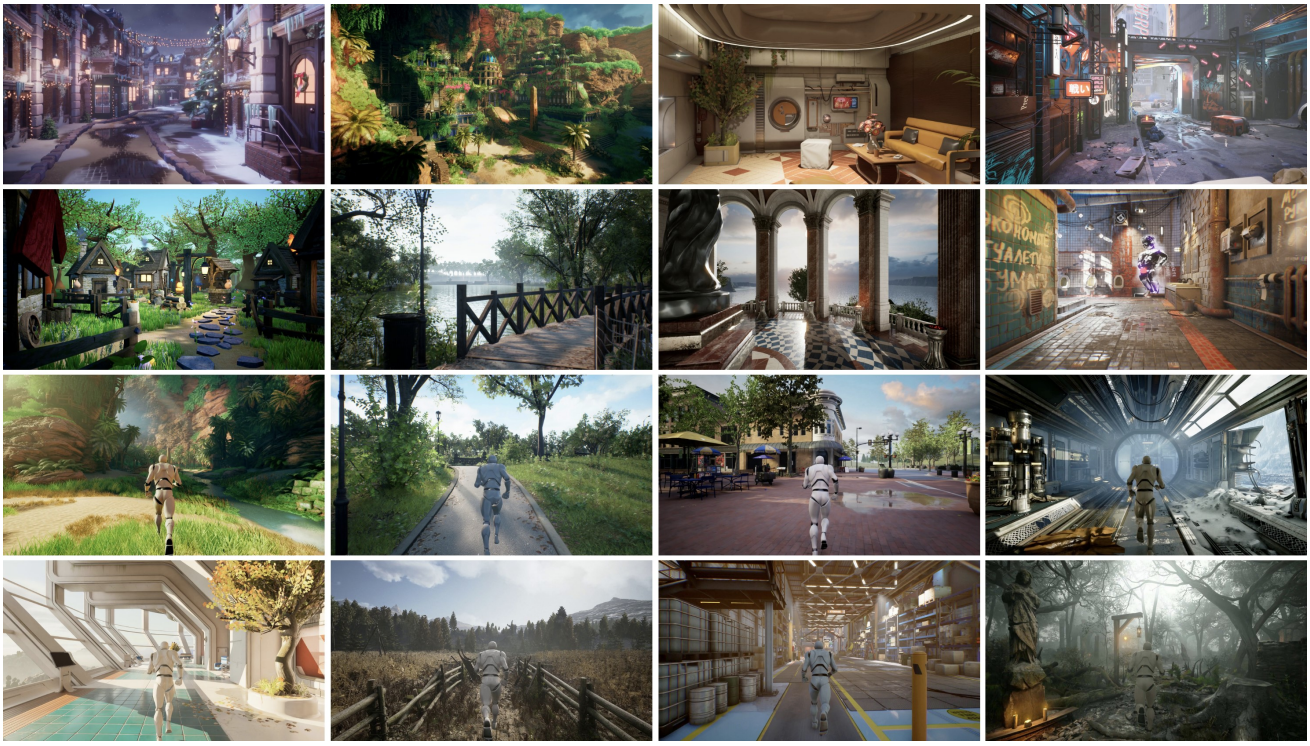


Fig. 2. Examples of 3D scenes from our ReVQ-2k dataset, featuring urban, interior, and landscape environments under different weather conditions.

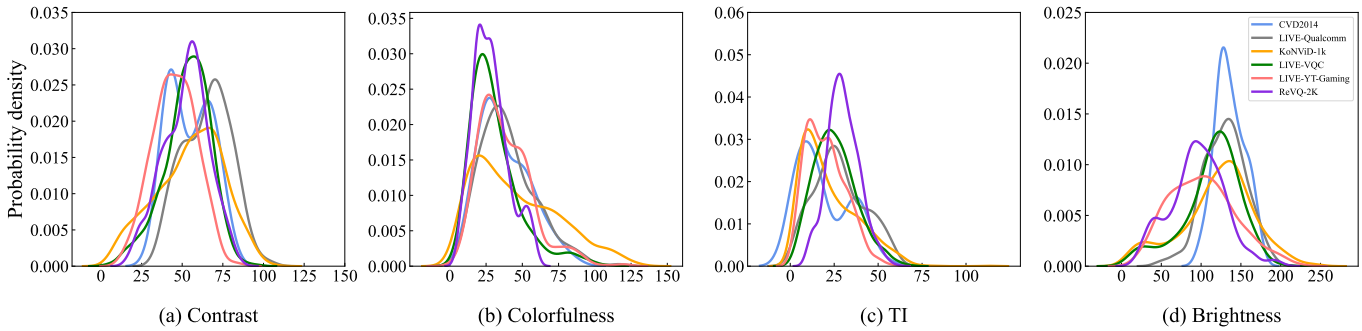



Fig. 3. Distributions of low-level quantitative attributes of our proposed ReVQ-2k dataset and five existing datasets.

## B. Subjective Quality Study

*1) Quality Scores: OA-MOS and TS-MOS:* To quantify the perceived video quality within the ReVQ-2k dataset, we employ the mean opinion score (MOS) in our user study, which offers an intuitive reflection of user-perceived quality. Our subjective quality assessment utilizes the overall mean opinion score (**OA-MOS**) for a comprehensive evaluation of rendered video quality. The OA-MOS, widely recognized in prior NR-VQA research [17], [18], [42], reflects the overall quality of videos by considering factors such as colorfulness, block artifacts, visual blurring, and temporal disruptions like flickering. Additionally, to address the specific challenges of assessing temporal stability of rendered videos, we introduce a new measure, temporal stability mean opinion score (**TS-MOS**). This measure evaluates the temporal stability of videos, focusing on issues like flickering, moving jaggies, and other temporal artifacts that significantly affect rendered video quality [29]. Incorporating TS-MOS into the evaluation adds extra oversight, allowing video quality prediction models to more accurately assess and predict the quality of rendered videos.

*2) Subjective Experiments:* In the subjective study, we implement a consistent stimulus evaluation process that allows participants to review the same video multiple times before rating it. The OA-MOS and TS-MOS are annotated on a scale of 1 to 5, with increments of 0.5, following the ITU-R absolute category rating scale [43], [44]. Quality scores range from 1 (”Bad”) to 5 (”Excellent”), with detailed criteria for these ratings provided in Tab. II.

The experiments for video quality rating are conducted with 17 trained annotators (10 male and 7 female with normal color vision), all of whom are familiar with rendered content. The research is conducted in a controlled laboratory setting, with each participant required to rate the entire video set on a specific type of display to ensure scoring consistency. The

TABLE II
DETAILED ANNOTATION CRITERIA FOR SUBJECTIVE VIDEO QUALITY SCORING.

| Score | TS-MOS Criteria | OA-MOS Criteria |
|---|---|---|
| 1 Bad | Significant temporal instabilities, flickering, or severe ghosting disrupt content continuity. | Difficulty to follow due to pronounced noise, block artifacts, Moiré patterns, blurring, substantial flickering, and lag. |
| 2 Poor | Noticeable temporal inconsistencies, flickering, or ghosting, but continuity is largely preserved. | Primary content is recognizable but degraded by considerable noise, block artifacts, blurring, noticeable flickering, and aliasing. |
| 3 Fair | Minor temporal artifacts; non-severe flickering or ghosting does not majorly impact continuity. | Primary content is clear with some distortions like mild noise, non-severe flickering, visual blurring, or noticeable false edges. |
| 4 Good | Temporal artifacts present but not distracting; continuity well-maintained with minimal flickering. | Clear primary subject with negligible noise or blurring and minor textural or edge distortions, not significantly affecting the experience. |
| 5 Excellent | Excellent temporal stability; no noticeable false edges, flickering, or ghosting. | Primary subject depicted with exceptional clarity, free of distracting distortions, and showing high-quality textural details. |

display monitor is color-calibrated to the sRGB standard, with brightness adjusted to 200 cd/$\mathrm{m}^2$ and the white point set to 6500K. The display's refresh rate is adjusted to 60 Hz to match the frame rate of the videos. Desktop monitors are positioned to ensure a comfortable viewing distance of 2 to 3 feet, akin to the experimental setup in [18], while the smartphone screen is positioned between 1 and 1.5 feet from the viewer. Three displays are used: a 27" AOC Q27P1U 2K IPS monitor, a 23.8" Dell P2422H 1080P IPS monitor, and a 6.7" AMOLED screen of an OPPO Find X3 Pro smartphone. To facilitate rapid annotation, custom software for both desktop and smartphone platforms has been developed to automate video playback and score recording. These tools can be downloaded from our project homepage.

Before the rating process begins, each participant views an instructional video that details the MOS measures and presents standard examples to illustrate the scoring criteria. During the training session, we specifically emphasize variations in temporal stability to ensure that participants fully understand the concepts of OA-MOS and TS-MOS scoring. Gold standard questions [9] are used to verify annotator performance. Annotators evaluate 10 'golden' videos, and those whose scores significantly diverge from the established standards (a difference greater than 1) are excluded from further stages of the study. All annotators in our tests successfully complete the training sessions, meeting the qualification criteria.

The subjective quality study is divided into three sessions: 720p on a smartphone screen, 1080p on a desktop monitor, and 2K on a desktop monitor. To prevent annotator fatigue, the test process is organized into rounds with rest periods; each round involves evaluating approximately 200 videos over 25 minutes, followed by a 10-minute rest period. Participants may opt to partake in one or more sessions, with compensation provided accordingly. At the conclusion of the subjective experiments, a total of 59,910 video quality ratings are collected from the 17 trained annotators.

*3) MOS Annotation Analysis:* **Data cleaning:** To ensure the validity of the MOS annotations, we implement three data cleaning methods in accordance with ITU-R BT.500-14 [43]: 1) We reject participants who have at least one annotation score on gold standard videos that deviates by more than 1 unit, as discussed in Sec. III-B2. 2) We include reappearing videos in the annotation process; participants are rejected if the difference in their ratings for any of these videos exceeds 1 unit. 3) We calculate correlation metrics, the Pearson linear correlation coefficient (PLCC) and the Spearman rank-order correlation coefficient (SRCC), between a participant's annotations and the average scores of other participants. Annotations from any participant whose correlation falls below 0.8 are rejected. In our experiments, no participants are excluded based on constraints 1) and 2); one participant is excluded due to failing to meet the criteria in constraint 3). From 16 qualified subjects, we have collected 57,450 valid ratings. The averages of the cleaned annotations serve as the MOS results for the videos.

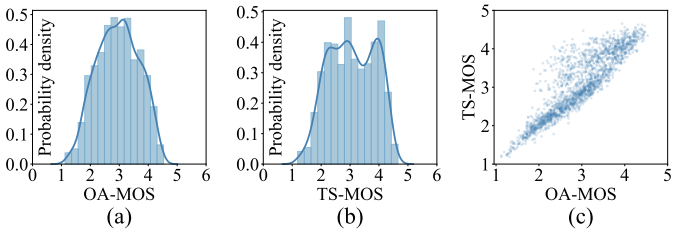



Fig. 4. Histograms (a) and (b) and scatter plot (c) of OA-MOS and TS-MOS from our proposed ReVQ-2k dataset. The project homepage includes separate data analyses for each resolution and display.

**Results distribution:** Fig. 4 shows the distribution of TS-MOS and OA-MOS scores. The kurtosis values [45] for the two distributions in Fig. 4 (a) and (b) are 2.231 and 1.955, respectively. Distributions with kurtosis values below 3 exhibit a plateau shape, indicating a more uniform distribution that captures quality levels across the video dataset effectively, thereby providing a robust basis for algorithm test [17]. Fig. 4 (c) shows scatter plots of OA-MOS and TS-MOS, revealing several critical insights. First, a large number of videos show a positive correlation between TS-MOS and OA-MOS, implying that videos with good temporal stability have higher overall quality. Second, some data points with low OA-MOS and high TS-MOS suggest that, while these videos maintain good temporal stability, they may suffer from image issues such as blurring or insufficient exposure, lowering their overall quality. Third, there are no data points in the quadrants for high OA-MOS and low TS-MOS. This is primarily because some videos with high image quality rarely exhibit extremely poor temporal stability, whereas some videos with poor temporal stability

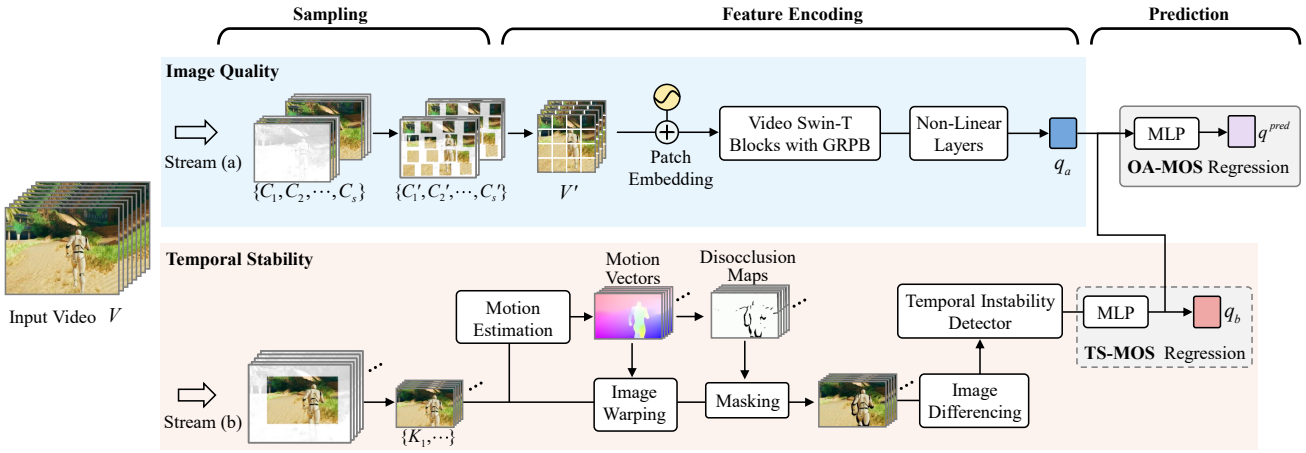



Fig. 5. Overview of our NR-VQA model. Given an input video $V$, our model employs two streams, (a) and (b), to extract features relevant to image quality and temporal stability, respectively. Stream (a) utilizes the FAST-VQA framework [13] to normalize input data and feeds the sampled video $V'$ into a Swin-T model, which extracts features that are processed through nonlinear layers to derive the image quality score $q_a$. Stream (b) assesses temporal stability by aligning frames using motion estimation, image warping, and occlusion removal, followed by applying image differencing to compute the temporal score $q_b$. Both scores, $q_a$ and $q_b$, are integrated using an MLP to predict the overall video quality. Notably, stream (b) can be trained using TS-MOS labels to improve the accuracy of NR-VQA for rendered videos.

rarely achieve high perceived quality ratings. These findings emphasize the importance of temporal stability in assessing the quality of rendered videos.

## IV. Our NR-VQA Method

### A. Overview

Building on the ReVQ-2k dataset, we develop a new two-stream NR-VQA metric to predict rendered video quality. Fig. 5 illustrates the architecture of our proposed model, which is divided into two main components: image quality assessment and temporal stability analysis. In the image quality assessment stream, we evaluate the overall image quality of videos. Given the extensive analysis in existing literature, we adopt the cropping strategy and Swin transformer (Swin-T) model employed in FAST-VQA [13] (see stream (a) in Fig. 5). This stream considers factors such as clarity and appropriate exposure, in alignment with existing NR-VQA methods [6], [10], and evaluates static rendering artifacts like Moiré patterns. In the temporal stability analysis stream (see stream (b) of Fig. 5), we crop a series of images from consecutive video frames, align them via motion estimation, and assess their temporal stability using image differencing. The results from both streams are then combined through a multilayer perceptron (MLP) to regress the overall video quality, integrating insights from Sec. III-B3.

It is worth noting that the entire model can be trained using only the OA-MOS; however, the TS-MOS can be used to train stream (b) prior to the entire model training. An ablation study presented in Sec. V-E1 shows that using temporal stability evaluations as additional supervision consistently improves prediction accuracy.

### B. Stream (a): Image Quality Evaluation

For the evaluation of overall image quality (specifically assessing issues such as blurring, noise, overexposure, and static rendering artifacts), extensive research [7], [10], [26] has been conducted in the field of NR-VQA for camera-captured videos. Considering the minor gaps in assessing these aspects between rendered and camera-captured videos, we directly employ an existing SOTA NR-VQA practice, FAST-VQA [13], for stream (a) of our method. The FAST-VQA method is selected for its well-designed architecture, which has proven to provide efficient and effective assessments of video quality.

As shown in Fig. 5, our input video $V = \{F_1, F_2, \ldots, F_z\}$ consists of $z$ frames, with $F_i$ denoting the $i$-th frame. In stream (a), the video is segmented into $s$ clips, each containing $t = z/s$ consecutive frames. For each clip $C_i = \{F_{i \cdot t}, F_{i \cdot t+1}, \ldots, F_{i \cdot t+t-1}\}$, we randomly select $m$ consecutive frames to form a subset $C_i' = \{F_j, F_{j+1}, \ldots, F_{j+m-1}\}$, where $j$ is randomly chosen from $(i \cdot t)$ to $(i \cdot t + t - m)$. Each frame $F_j$ in $C_i'$ is then divided into an $n \times n$ grid; then, a $k \times k$ image patch is cropped from each grid cell. The cropped patches are concatenated to form an image $F_i'$ of dimensions $(n \cdot k) \times (n \cdot k)$. This procedure is replicated for all images in image subsets $\{C_1', C_2', \ldots, C_s'\}$, producing a resampled video $V'$. This approach utilizes key parameters: $s = 8$, $m = 4$, $n = 7$, and $k = 32$. Such a sampling strategy not only significantly reduces data volume by eliminating potential redundancy but also standardizes input videos of varying lengths and resolutions into a uniform format without the need for resizing. Readers are referred to FAST-VQA [13] for further details.

After the sampling process, the video $V'$ undergoes patch embedding and is processed using a Swin-T model to extract high-level features related to video quality assessment. It is important to note that $V'$ is constructed from small patches, which can introduce discontinuities at the seams of patches. Consequently, the feature extraction network must be carefully designed to avoid misinterpreting these seams as image artifacts. To address this, FAST-VQA [13] employs non-overlapping pooling kernels, effectively preventing the

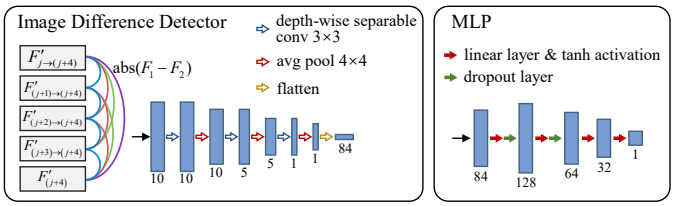



Fig. 6. Detailed structures of the image difference detector and the MLP model. The numbers in the diagram represent the feature channels.

misinterpretation of patch seams as artifacts. Additionally, FAST-VQA introduces gated relative position biases (GRPB) within the Swin-T to enhance the performance of the self-attention layers. Finally, the features extracted from each patch are processed through non-linear layers to derive a quality score $q_a$, which assesses the image quality of the input video.

### C. Stream (b): Temporal Stability Evaluation

While some existing NR-VQA methods have started to incorporate temporal stability into video quality assessment, they often lack the capability to effectively detect temporal artifacts such as flickering and moving jaggies, which are rarely present in camera-captured content. However, in rendered videos, temporal artifacts are prevalent and significantly impact video quality [35]. Although FAST-VQA employs temporally contiguous image patches in video sampling, it fails to detect instability due to the common misalignment of moving objects in videos. To address this, we propose a temporal artifacts detector that forms stream (b) of our metric. This stream uses motion estimation to align objects across frames and utilizes image differencing to detect pixel-level temporal instability, thereby playing a vital role in the assessment of rendered video quality.

As depicted in stream (b) of Fig. 5, to reduce computational overhead and standardize input data shape, we first sample the input video into a normalized shape. We uniformly extract 10 subsets from the input video $V$, each containing 5 consecutive frames. To normalize the resolution, these subsets are cropped to a uniform resolution of $h = 480, w = 800$, producing standardized subsets $\{K_1, K_2, \ldots, K_{10}\}$. The cropping location is consistent across all frames within each subset but is randomly selected relative to the original frame positions. Each subset $K_i$ then undergoes motion estimation, generating motion vectors $\{M_j, M_{j+1}, M_{j+2}, M_{j+3}\}$ for frames $\{F'_j, F'_{j+1}, F'_{j+2}, F'_{j+3}\}$ to $F'_{j+4}$, where $F'_j$ represents the cropped image of frame $F_j$. We employ the SOTA motion estimation algorithm, dense optical tracking (DOT) [46], to efficiently track points across video frames by capturing key motion tracks from dynamic boundaries. Additionally, we have enhanced the DOT algorithm to mark occluded areas, enabling the concurrent generation of disocclusion maps $\{D_j, D_{j+1}, D_{j+2}, D_{j+3}\}$. These maps identify regions in $F'_{j+4}$ where objects in $\{F'_j, F'_{j+1}, F'_{j+2}, F'_{j+3}\}$ are occluded, effectively marking pixels lacking temporal correspondences. After generating motion vectors and disocclusion maps, each subset $K_i$ is subjected to backward warping and disoccluded pixel removal, resulting in an aligned subset $K'_i = \{F'_{j\to(j+4)}, F'_{(j+1)\to(j+4)}, \ldots, F'_{j+4}\}$, with $F'_{j\to(j+4)}$ obtained by:

$$F'_{j\to(j+4)} = \text{Warping}(F_j, M_j) \cdot D, \tag{1}$$

where $D$ represents the overlap of the four disocclusion maps. Then, the processed image subsets $\{K'_1, K'_2, \ldots, K'_{10}\}$ undergo image differencing over various time spans. As shown in Fig. 6, we calculate image differencing for adjacent frames and for frames separated by one, two, and three intervals, enabling detection of flickering across various frequencies. The image differences are then fed into a depth-wise separable convolutions-based [47] image difference detector to evaluate pixel-level temporal stability. Finally, the stability maps from all subsets are subjected to average pooling and an MLP to regress the temporal stability score $q_b$. The detailed structures of the image difference detector and the MLP are presented in Fig. 6.

### D. Final MOS Prediction

After obtaining the image quality score $q_a$ (which can be seen as the result of the FAST-VQA [13] method) and temporal stability score $q_b$, we aim to determine their relationship to the overall video quality, expressed as $q^{pred} = f(q_a, q_b)$. As discussed in Sec. III-B3, there exists a non-linear relationship between the temporal stability of rendered videos and their overall perceived quality. Here, we implement an MLP to model the complex non-linear mapping between ($q_a$, $q_b$) and $q^{pred}$. This MLP utilizes a configuration similar to that used in stream (b), but with only a quarter of the number of neurons to prevent overfitting. Experimental results presented in Sec. V-E1 also demonstrate that this non-linear mapping approach achieves superior accuracy compared to linear mapping methods.

For the loss functions, extensive exploration has been conducted in prior studies [8], [9], with research primarily focusing on the correlation between predicted scores and ground truth. Consistent with prevalent practices, we adopt the widely used PLCC and ranking loss functions. Given a batch of predicted quality scores $Q^{pred} = \{q_1^{pred}, q_2^{pred}, \ldots, q_s^{pred}\}$ and ground truth labels $Q = \{q_1, q_2, \ldots, q_s\}$, these loss functions are defined as:

$$\begin{aligned} L_{\text{PLCC}} &= \left(1 - \frac{\sum_{i=1}^{s}\left(q_i^{pred} - a\right)(q_i - b)}{\sqrt{\sum_{i=1}^{s}(q_i^{pred} - a)^2 \sum_{i=1}^{s}(q_i - b)^2}}\right)/2, \\ L_{\text{ranking}} &= \frac{1}{s^2}\sum_{i=1}^{s}\sum_{j=1}^{s}\max\left((q_j^{pred} - q_i^{pred})\text{sgn}(q_i - q_j), 0\right), \end{aligned} \tag{2}$$

where $a$ and $b$ are the mean values of $Q^{pred}$ and $Q$, respectively, and $\text{sgn}(\cdot)$ denotes the sign function. We empirically set a weight $\alpha$ to 0.3 and combine the loss functions as:

$$\text{LOSS} = L_{\text{PLCC}} + \alpha \cdot L_{\text{ranking}}, \tag{3}$$

which is used to train both stream (b) and our entire model.

TABLE III
QUANTITATIVE COMPARISON OF NR-VQA METHODS ON THE REVQ-2K DATASET (720P, 1080P, AND 2K RESOLUTIONS). METRICS SHOWN INCLUDE PLCC AND SRCC. **BOLDFACE** DENOTES THE BEST PERFORMANCE AND UNDERLINE INDICATES THE SECOND-BEST PERFORMANCE FOR EACH METRIC.

| Methods | ReVQ-2k (720p) | | ReVQ-2k (1080p) | | ReVQ-2k (2k) | | Weighted Average | |
|---|---|---|---|---|---|---|---|---|
| | SRCC↑ | PLCC↑ | SRCC↑ | PLCC↑ | SRCC↑ | PLCC↑ | SRCC↑ | PLCC↑ |
| VSFA [10] | 0.785 | 0.767 | 0.755 | 0.738 | 0.772 | 0.780 | 0.771 | 0.762 |
| CNN-TLVQM [26] | 0.782 | 0.784 | 0.773 | 0.774 | 0.762 | 0.779 | 0.772 | 0.779 |
| GSTVQA [6] | 0.839 | 0.829 | 0.836 | 0.817 | 0.826 | 0.822 | 0.834 | 0.823 |
| SimpleVQA [8] | 0.801 | 0.808 | 0.759 | 0.785 | 0.780 | 0.796 | 0.780 | 0.796 |
| DOVER [9] | 0.821 | 0.829 | 0.809 | 0.817 | 0.812 | 0.815 | 0.814 | 0.820 |
| FAST-VQA [13] | 0.846 | 0.845 | 0.852 | 0.856 | 0.823 | 0.849 | 0.840 | 0.850 |
| MBVQA [48] | 0.818 | 0.826 | 0.803 | 0.821 | 0.814 | 0.839 | 0.812 | 0.829 |
| Ours- | 0.854 | 0.859 | 0.864 | 0.863 | 0.836 | 0.853 | 0.851 | 0.858 |
| Ours | **0.882** | **0.884** | **0.885** | **0.887** | **0.869** | **0.874** | **0.879** | **0.882** |

## V. EXPERIMENTS

This section describes how we implemented our method and compares it to existing VQA methods on both the proposed ReVQ-2k dataset and the existing NR-VQA dataset. Then we perform ablation studies to determine the impact of individual components in our model.

### A. Implementation Details

In our method, stream (b) can be independently trained using TS-MOS labels or integrated with the entire model trained using only OA-MOS labels. We implement both strategies using a batch size of 16 and the Adam optimizer with a learning rate of 0.001. The remaining training configurations are maintained consistent with those used in the FAST-VQA method. Given that the motion estimation phase is typically more time-consuming compared to neural network inference, we recommend precomputing motion vectors for the dataset to reduce both training and test times. Our experiments are conducted on a desktop PC equipped with an NVIDIA GeForce RTX 4090 GPU, an Intel i7-13700K CPU, and 64GB of memory.

### B. Evaluation Setups

**Datasets.** To evaluate the performance of NR-VQA methods, we utilize the proposed ReVQ-2k dataset, along with established NR-VQA datasets[1], including CVD2014 [17], LIVE-Qualcomm [18], KoNViD-1k [4], LIVE-VQC [5], and LIVE-YT-Gaming [19]. Detailed descriptions of these datasets are presented in Tab. I. The ReVQ-2k dataset includes videos at three resolutions: 2K, 1080P, and 720P, each played and annotated on their respective displays. Therefore, comparisons on the ReVQ-2k dataset are performed separately for each resolution. Note that we consider videos from each 3D scene as unique entities for the training and test phases, ensuring scene independence during the training/test set splitting. During the test, results from each scene are independently calculated and then combined to produce a weighted average. This approach addresses potential biases due to varying scene content, more closely reflecting real-world applications.

**Evaluation Metrics.** We assess NR-VQA models using two commonly used criteria: the PLCC and the SRCC. The PLCC is employed to evaluate the accuracy of predictions by measuring the linear relationship between predicted and actual values. The SRCC is used to assess the monotonicity of predictions, evaluating how consistently the predictions preserve the ordinality of the actual values. When addressing datasets with varying MOS scales, we utilize a logistic function $g$, as suggested by Li *et al*. [10], to map the predicted scores $o$ to the corresponding subjective scores $s$:

$$g(o) = \frac{\beta_1 - \beta_2}{1 + e^{-\frac{o-\beta_3}{\beta_4}}} + \beta_2, \tag{4}$$

where $\beta_1 = \max(s)$, $\beta_2 = \min(s)$, $\beta_3 = \text{mean}(o)$, and $\beta_4 = \text{std}(o)/4$.

### C. Comparisons on Rendered Video Dataset

We compare our method against existing SOTA baselines, including VSFA [10], CNN-TLVQM [26], GSTVQA [6], SimpleVQA [8], DOVER [9], FAST-VQA [13], and the full version of MBVQA [48]. To assess their performance on the ReVQ-2k dataset, we utilize the source code provided by the authors, training their models under the recommended settings to ensure optimal performance. All methods, including ours, are initially pretrained on the training set of the large-scale LSVQ dataset [16], and then fine-tuned on the test datasets. We implement our approach in two variants: one trained solely with OA-MOS labels, denoted as "Ours-", and another that incorporates TS-MOS labels for additional supervision, denoted as "Ours". To minimize variability due to random dataset partitioning, we repeatedly split the 3D scenes into training and test sets five times, maintaining a splitting ratio of approximately 8:2 for training and test. The specifics of these splits are included in the public release of the dataset.

The quantitative analysis in Tab. III clearly demonstrates that our proposed model significantly outperforms existing SOTA methods. Without temporal stability supervision, our model (Ours-) exceeds the top-performing baseline by 1.1%

[1]We did not conduct experiments on the largest NR-VQA dataset, LSVQ [16], for two main reasons. First, the dataset's extensive video collection would demand an impractically large amount of time for motion estimation of our method. Second, the dataset primarily consists of camera-captured videos, which are not directly relevant for evaluating the performance of our rendering-oriented method.


Fig. 7. Visual comparison of video quality predictions by FAST-VQA and our method on the ReVQ-2k dataset. Temporal profiles (orange boxes) are depicted through a column of pixels across temporal frames. $q_{FAST}$ represents predictions from FAST-VQA, while $q_b$, $q^{pred}$, and $q^{GT}$ denote the results from stream (b), our final result, and the ground truth video quality (OA-MOS), respectively. The predictions are rescaled using the mean and standard deviation of the ground truth annotations. This figure illustrates how our method integrates image quality (using $q_{FAST}$ as a reference) with temporal stability assessments ($q_b$) to achieve a comprehensive evaluation of rendered video quality ($q^{pred}$) .

TABLE IV
COMPARISON OF NR-VQA METHODS ON ESTABLISHED DATASETS INCLUDING CVD2014, LIVE-QUALCOMM, KONVID-1K, LIVE-VQC, AND LIVE-YT-GAMING.

| Methods | CVD2014 | | LIVE-Qualcomm | | KoNViD-1k | | LIVE-VQC | | LIVE-YT-Gaming | |
|---|---|---|---|---|---|---|---|---|---|---|
| | SRCC↑ | PLCC↑ | SRCC↑ | PLCC↑ | SRCC↑ | PLCC↑ | SRCC↑ | PLCC↑ | SRCC↑ | PLCC↑ |
| VSFA [10] | 0.850 | 0.869 | 0.708 | 0.774 | 0.794 | 0.799 | 0.718 | 0.771 | 0.784 | 0.819 |
| CNN-TLVQM [26] | 0.863 | 0.880 | 0.810 | 0.833 | 0.816 | 0.818 | 0.825 | 0.834 | 0.855 | 0.866 |
| GSTVQA [6] | 0.831 | 0.844 | 0.801 | 0.825 | 0.814 | 0.825 | 0.788 | 0.796 | 0.850 | 0.860 |
| SimpleVQA [8] | 0.834 | 0.864 | 0.722 | 0.774 | 0.792 | 0.798 | 0.740 | 0.775 | 0.814 | 0.836 |
| DOVER [9] | 0.858 | 0.881 | 0.736 | 0.789 | 0.892 | 0.900 | 0.853 | 0.872 | 0.882 | **0.906** |
| FAST-VQA [13] | **0.883** | **0.901** | 0.807 | 0.814 | 0.893 | 0.887 | 0.853 | 0.873 | 0.869 | 0.880 |
| MBVQA [48] | **0.883** | **0.901** | **0.832** | **0.842** | **0.901** | **0.905** | **0.860** | **0.880** | 0.867 | 0.902 |
| Ours- | 0.881 | 0.895 | 0.816 | 0.828 | 0.896 | 0.894 | 0.857 | 0.869 | **0.891** | 0.904 |

and 0.8% in PLCC and SRCC, respectively. This margin increases to 3.9% and 3.2% when our model (Ours) incorporates temporal stability scores for training. While the DOVER method [9], which incorporates aesthetic opinions, proves effective for camera-captured videos, using a pretrained aesthetic evaluation module offers no improvement in NR-VQA performance on rendered datasets. Similarly, the MB-VQA method [48], introducing spatial and temporal rectifiers, also fails to deliver advantages for rendered data.

Fig. 7 illustrates the visual results with temporal profiles of a column of pixels across temporal frames, effectively visualizing the temporal stability of videos. Notably, since FAST-VQA does not directly evaluate video temporal stability, there are frequent discrepancies between its predictions ($q_{FAST}$) and the ground truth ($q^{GT}$). For instance, in panels (a), (d), and (g), where videos exhibit good temporal stability, FAST-VQA often underestimates the overall quality; conversely, in panels (b), (c), (e), and (f), which are characterized by poor temporal stability, it yields overly high quality estimates. By integrating the temporal stability evaluation $q_b$, we enhance the assessment of temporal artifacts, thus aligning the results $q^{pred}$ more closely with human-annotated scores. This analysis demonstrates the effectiveness of incorporating temporal stability assessments in NR-VQA for rendered videos.

### D. Comparisons on Existing NR-VQA Datasets

We also test our method on existing datasets of camera-captured content, despite these tests are not directly related to the evaluation of our metric designed for rendered videos. For the CVD2014 [17], LIVE-Qualcomm [18], KoNViD-1k [4], LIVE-VQC [5], and LIVE-YT-Gaming [19] datasets, we adhere to established research protocols by conducting experiments using 10 random train-test splits, allocating 80% of the data for training and 20% for test. All methods are pretrained on the training set of the large-scale LSVQ dataset [16]. Because these datasets lack TS-MOS annotations, we

only present the "Ours-" model, which is trained with only OA-MOS labels. The comparative results, shown in Tab. V-B, demonstrate that our method is competitive in assessing camera-captured video quality. For most datasets, our method achieves top-2 performance in both PLCC and SRCC scores. Although the improvements compared to baselines on these datasets are not as pronounced as those observed on the rendered video dataset, our model still exhibits strong robustness. Specifically, our method shows superior performance on the LIVE-YT-Gaming dataset, showing potential future extensions to NR-VQA for cloud gaming and streaming games.

### E. Ablation Study

*1) Effectiveness of Temporal Stability Evaluation:* We evaluate the contributions of streams (b), TS-MOS annotations, and non-linear MOS regression within our model through seven variants: 1) using only stream (a); 2) using only stream (b), trained and evaluated on OA-MOS labels; 3) using only stream (b), trained on TS-MOS labels but evaluated on OA-MOS labels; 4) combining streams (a) and (b) with a linear function; 5) combining streams (a) and (b) with a non-linear function; 6) incorporating TS-MOS training into variant 4; and 7) incorporating TS-MOS training into variant 5. The tests are carried out on the ReVQ-2k dataset (720p and 1080p) using five train-test splits. The results, detailed in Tab. V, demonstrate significant enhancements in video quality prediction accuracy with the integration of temporal stability evaluation by streams (b) and the TS-MOS annotations. Including stream (b) and using TS-MOS labels for training results in average improvements of 4.5% in SRCC and 4.0% in PLCC, and using non-linear for MOS regression further enhances the results.

*2) Effectiveness of Motion Estimation and Masking:* To effectively assess temporal stability in videos, our method integrates motion estimation and occlusion masking in stream (b). We conduct experiments to validate the effectiveness of these components by comparing three model configurations: one without motion estimation and masking, one with motion estimation but no masking, and our final model with both.[2] The results, reported in Tab. VI, reveal that the absence of motion estimation significantly compromises the model's accuracy in predicting temporal stability, with an SRCC of 0.170 and a PLCC of 0.163. Introducing motion estimation alone improves these metrics to 0.782 and 0.788, respectively. Our complete model, incorporating both motion estimation and masking, achieves further improvements, attaining an SRCC of 0.842 and a PLCC of 0.853. These findings validate the effectiveness of incorporating motion estimation and masking for enhanced temporal stability assessment.

TABLE V
COMPARISON OF VIDEO QUALITY PREDICTION ACCURACY WITH STREAMS (B), TS-MOS LABELS, AND THE MLP MODEL ON THE REVQ-2K DATASET.

| Methods | ReVQ-2k (720p) | | ReVQ-2k (1080p) | |
|---|---|---|---|---|
| | SRCC↑ | PLCC↑ | SRCC↑ | PLCC↑ |
| $q_a$ | 0.843 | 0.840 | 0.825 | 0.836 |
| $q_b$ | 0.728 | 0.713 | 0.753 | 0.768 |
| $q_b$ & TS-MOS | 0.683 | 0.695 | 0.721 | 0.729 |
| Lin$(q_a, q_b)$ | 0.848 | 0.858 | 0.859 | 0.861 |
| MLP$(q_a, q_b)$ | 0.854 | 0.859 | 0.864 | 0.863 |
| Lin$(q_a, q_b)$ & TS-MOS | 0.876 | 0.880 | 0.881 | 0.876 |
| MLP$(q_a, q_b)$ & TS-MOS | **0.882** | **0.884** | **0.885** | **0.887** |

TABLE VI
COMPARISON OF MOTION ESTIMATION AND MASKING ON TEMPORAL STABILITY PREDICTION, WITH RESULTS EVALUATED BY TS-MOS LABELS.

| Methods | ReVQ-2k (720p) | | ReVQ-2k (1080p) | |
|---|---|---|---|---|
| | SRCC↑ | PLCC↑ | SRCC↑ | PLCC↑ |
| ***w/o*** motion & masking | 0.187 | 0.214 | 0.152 | 0.113 |
| ***w*** motion, ***w/o*** masking | 0.779 | 0.786 | 0.785 | 0.790 |
| ***w*** motion & masking | **0.831** | **0.856** | **0.852** | **0.850** |

## VI. APPLICATIONS

This section presents two real-world applications of our NR-VQA metric. First, we apply the metric to assess video quality across various closed-source supersampling methods for mobile real-time rendering, addressing scenarios where videos cannot be perfectly aligned. Second, we utilize the metric to evaluate the perceived quality of various frame generation strategies for real-time rendering. These applications demonstrate the practical utility of our NR-VQA metric in rendering program development.

### A. Benchmarking Mobile Supersampling Methods

In the real-time rendering, supersampling is extensively employed to achieve anti-aliasing or super-resolution (SR) in images, with its application rapidly expanding in mobile real-time rendering. Lightweight supersampling methods suitable for mobile platforms include hardware-dependent solutions such as Snapdragon's game SR [49] and Pixelworks' hardware SR method [50], as well as hardware-independent methods like AMD FSR 1.0 [51] and 2.0 [38], and MNSS [52]. Additionally, our team has recently developed a new lightweight SR method [53]. However, comparing our SR method to existing ones is challenging due to misalignment issues in videos produced by various closed-source SR algorithms, where object positions and character poses cannot be consistently aligned across videos. This misalignment makes full-reference methods such as SSIM [2] and VMAF [54] impractical. In constrast, our NR-VQA method enables effective quality evaluation of these non-aligned videos.

To evaluate the performance of different SR methods, we collect videos from two 3D game scenes, each with $\times 2$ resolution upscaling. Each method generates five 8-second videos at 60 FPS / 720p for each scene. To minimize quality biases caused by content variations, efforts are made to ensure that the video content from different SR methods is as similar as possible. We employ our NR-VQA metric, calibrated on the ReVQ-2k (720p) dataset, to perform perceptual video quality scoring. For commercial considerations, we anonymize

[2]Note that only stream (b) is tested, with TS-MOS labels used for both training and test.

TABLE VII
QUALITY ASSESSMENT OF VARIOUS SUPERSAMPLING METHODS, PRESENTING VIDEO QUALITY SCORES (RESCALED USING THE LABELED OA-MOS) FROM OUR NR-VQA METRIC AND HUMAN-LABELED OA-MOS ACROSS TWO SCENES.

| Methods | Scene 1 | | Scene 2 | |
|---|---|---|---|---|
| | Ours ↑ | OA-MOS ↑ | Ours ↑ | OA-MOS ↑ |
| $A_1$ | 1.72 | 1.65 | 1.89 | 1.82 |
| $A_2$ | 1.88 | 1.79 | 2.14 | 1.89 |
| $A_3$ | 2.89 | 2.85 | 2.92 | 3.18 |
| $A_4$ | **3.12** | **3.71** | **3.45** | **3.41** |
| $O_1$ | 2.52 | 2.33 | 2.54 | 2.51 |
| $O_2$ | 2.71 | 2.52 | 2.74 | 2.79 |
| $O_3$ | 2.99 | 3.02 | 3.01 | 3.19 |

TABLE VIII
QUALITY ASSESSMENT OF FRAME GENERATION STRATEGIES, SHOWING SCORES FROM OUR NR-VQA METRIC (RESCALED) AND HUMAN-LABELED OA-MOS.

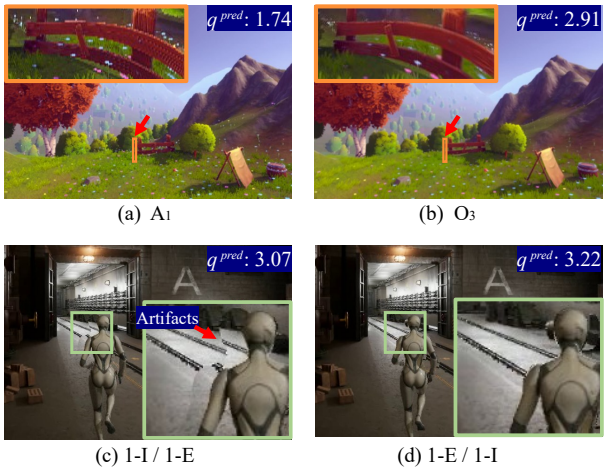
| Strategies | Scene 3 | | Scene 4 | |
|---|---|---|---|---|
| | Ours ↑ | OA-MOS ↑ | Ours ↑ | OA-MOS ↑ |
| A-R | 3.92 | 4.15 | 3.70 | 3.94 |
| 1-I | 3.63 | 3.71 | 3.41 | 3.48 |
| 1-E | 3.46 | 3.44 | 3.23 | 3.07 |
| 2-I | 3.64 | 3.65 | 3.28 | 3.35 |
| 1-E/1-I | 3.31 | 3.43 | 3.15 | 3.12 |
| 1-I/1-E | 3.08 | 2.91 | 2.96 | 2.84 |
| 2-E | 3.05 | 2.82 | 2.87 | 2.81 |

the names of the SR methods, using codes $O_1, O_2, O_3$ for versions of our SR model, and $A_1, A_2, A_3, A_4$ for comparative methods. To ensure a fair and consistent evaluation, all videos are annotated by the trained participants from the proposed ReVQ-2k dataset to derive the reference OA-MOS scores.

The experiment results, as reported in Tab. VII, indicate that method $A_4$ achieves the highest quality scores across all scenes. The results for $O_3$ are comparable to those of $A_3$, while $A_1$ and $A_2$ underperform. Additionally, the strong correlation between our model's predicted video quality and the manually annotated OA-MOS confirms the accuracy of our automated method. The proposed NR-VQA metric facilitates immediate quantitative evaluations of our SR algorithm and other closed-source methods, significantly enhancing efficiency and reducing labor costs. As illustrated in Fig. 8 (a) and (b), high-quality rendered videos effectively capture object details within the scenes and exhibit fewer aliasing artifacts. In contrast, method $A_1$ suffers from noticeable spatial aliasing and temporal instability, leading to less satisfactory outcomes. In this application, although method $A_4$ delivers superior results, its computational costs exceed our budget. The performance of $O_3$ is comparable to that of $A_3$, but it is achieved at a lower cost and within our budget. Using the proposed NR-VQA metric, we can rapidly and automatically validate the performance of the developed method.

### B. Evaluating Frame Generation Strategies

Frame generation techniques (interpolation and extrapolation) are used in real-time rendering applications to enhance frame rates. Mob-FGSR [53] introduces a method that can generate frames at arbitrary times, allowing for multiple interpolated and extrapolated frames. A significant challenge in deploying this algorithm is determining the optimal combination of interpolation and extrapolation to achieve the best video quality. Full-reference metrics are unsuitable since the estimated motion of generated frames typically do not align with reference images. Although user ratings provide a feasible method for quality assessment, they are labor-intensive and prone to inconsistency; as ratings from an individual can vary significantly from one day to the next. Thus, quantitative video quality assessments are crucial for efficient project development.



(a) A1 (b) O3
(c) 1-I / 1-E (d) 1-E / 1-I

Fig. 8. Examples of the applications. (a) and (b) show the supersampling results, temporal profiles (orange boxes), and predicted video quality scores $q^{pred}$. (c) and (d) illustrate the frame generation results and the predictions $q^{pred}$, with artifacts shown in light green boxes.

In this study, we utilize our NR-VQA metric to quantitatively assess various frame generation strategies. We collect five 8-second videos at 60 FPS / 1080p for each of two 3D game scenes using different strategies: all rendered frames ("A-R"), one interpolated frame ("1-I"), one extrapolated frame ("1-E"), two interpolated frames ("2-I"), two extrapolated frames ("2-E"), interpolation followed by extrapolation ("1-I/1-E"), and extrapolation followed by interpolation ("1-E/1-I"). For each strategy, we gather videos that are closely similar in content for each scene and employ annotators to provide reference OA-MOS labels. The videos are then assessed for quality using our NR-VQA metric. The results, shown in Tab. VIII, depict the video quality for each strategy. Across both scenes, "A-R" yields the highest quality scores, followed by "1-I" and "1-E", with "1-I" outperforming "1-E" due to having additional frame references. For strategies involving two frame generations, "2-I" exhibits the best performance, followed by "1-E/1-I" and then "1-I/1-E" (see Fig. 8 (c) and (d)), with "2-E" showing the lowest quality. These results align with our expectations. Given the latency issues associated with interpolation, we recommend strategies "1-E" or "1-E/1-I" for latency-sensitive games, and "1-I" or "2-I" for games where delays are less sensitive.

## VII. Limitations

Our NR-VQA dataset and metric for assessing rendered video quality have several limitations. First, the TS-MOS in our dataset primarily captures temporal instability but does not evaluate the fluidity or naturalness of object movements within the video. As a result, issues related to motion smoothness are not considered, which could be addressed in future work. Second, the NR-VQA metric shows varying sensitivity to different scene contents and motion velocities, potentially leading to biased scores in certain cases. To ensure a fair comparison, it is crucial that the scene contents and motion conditions of the compared videos are as consistent as possible. Third, stream (b) occasionally leads to less optimal outcomes. As seen in Fig. 7 panel (i), in cases where videos display good temporal stability but exhibit severe blurring, our method slightly overestimates the overall quality. To address this, future work could more precisely analyze the impact of temporal stability to enhance the accuracy of the metric. Finally, due to the use of motion estimation, the model's runtime is increased, typically taking over ten seconds to evaluate the quality of one video. This delay, however, remains acceptable for most practical applications.

## VIII. Conclusions

Accurate video quality evaluations are essential for numerous rendering applications, such as pipeline optimization and parameter selection. To address the challenges of assessing the quality of rendered videos that cannot be perfectly aligned and lack reference videos, we introduce a large rendering-oriented VQA dataset along with a novel NR-VQA metric specifically designed for rendered content. The dataset, termed ReVQ-2k, comprises 2,000 videos featuring a variety of 3D scenes and rendering settings, each annotated with overall quality labels (OA-MOS) and temporal stability labels (TS-MOS). Our NR-VQA metric provides a comprehensive evaluation of rendered videos by analyzing both overall image quality and temporal stability. Experiments on the ReVQ-2k dataset confirm the superior accuracy of our metric, consistently outperforming existing SOTA methods. Furthermore, we demonstrate the practical utility of our NR-VQA metric in two real-world applications, showing its capacity to reduce manual labor and accelerate the development of rendering algorithms. Our work establishes a robust benchmark and provides a baseline method for NR-VQA of rendered videos, offering valuable insights for related applications and future research.

## Acknowledgment

This work was supported by Key R&D Program of Zhejiang (No. 2024C01069), and the National Natural Science Foundation of China (Grant No. 62036010).

## References

[1] P. Andersson, J. Nilsson, T. Akenine-Möller, M. Oskarsson, K. Åström, and M. D. Fairchild, "FLIP: A difference evaluator for alternating images." *Proc. ACM Comput. Graph. Interact. Tech.*, vol. 3, no. 2, pp. 15–1, 2020.

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

[3] 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, pp. 1–19, 2021.

[4] V. Hosu, F. Hahn, M. Jenadeleh, H. Lin, H. Men, T. Szirányi, S. Li, and D. Saupe, "The konstanz natural video database (KoNViD-1k)," in *Ninth Int. Conf. Qual. Multimedia Exp.* IEEE, 2017, pp. 1–6.

[5] Z. Sinno and A. C. Bovik, "Large-scale study of perceptual video quality," *IEEE Trans. Image. Process.*, vol. 28, no. 2, pp. 612–627, 2018.

[6] B. Chen, L. Zhu, G. Li, F. Lu, H. Fan, and S. Wang, "Learning generalized spatial-temporal deep feature representation for no-reference video quality assessment," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 4, pp. 1903–1916, 2021.

[7] Y. Wang, J. Ke, H. Talebi, J. G. Yim, N. Birkbeck, B. Adsumilli, P. Milanfar, and F. Yang, "Rich features for perceptual quality assessment of UGC videos," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13 435–13 444.

[8] W. Sun, X. Min, W. Lu, and G. Zhai, "A deep learning based no-reference quality assessment model for ugc videos," in *Proc. ACM Int. Conf. Multimedia*, 2022, pp. 856–865.

[9] H. Wu, E. Zhang, L. Liao, C. Chen, J. Hou, A. Wang, W. Sun, Q. Yan, and W. Lin, "Exploring video quality assessment on user generated contents from aesthetic and technical perspectives," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 20 144–20 154.

[10] D. Li, T. Jiang, and M. Jiang, "Quality assessment of in-the-wild videos," in *Proc. ACM Int. Conf. Multimedia*, 2019, pp. 2351–2359.

[11] 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.

[12] J. L. Cardoso, B. Kerbl, L. Yang, Y. Uralsky, and M. Wimmer, "Training and predicting visual error for real-time applications," *Proc. ACM Comput. Graph. Interact. Tech.*, vol. 5, no. 1, pp. 1–17, 2022.

[13] H. Wu, C. Chen, L. Liao, J. Hou, W. Sun, Q. Yan, J. Gu, and W. Lin, "Neighbourhood representative sampling for efficient end-to-end video quality assessment," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 12, pp. 15 185–15 202, 2023.

[14] C. Keimel, A. Redl, and K. Diepold, "The TUM high definition video datasets," in *Fourth Int. Workshop Qual. Multimedia Exp.* IEEE, 2012, pp. 97–102.

[15] P. V. Vu and D. M. Chandler, "Vis3: An algorithm for video quality assessment via analysis of spatial and spatiotemporal slices," *J. Electron. Imaging*, vol. 23, no. 1, p. 013016, 2014.

[16] Z. Ying, M. Mandal, D. Ghadiyaram, and A. Bovik, "Patch-VQ:'patching up'the video quality problem," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14 019–14 029.

[17] M. Nuutinen, T. Virtanen, M. Vaahteranoksa, T. Vuori, P. Oittinen, and J. Häkkinen, "CVD2014—a database for evaluating no-reference video quality assessment algorithms," *IEEE Trans. Image. Process.*, vol. 25, no. 7, pp. 3073–3086, 2016.

[18] D. Ghadiyaram, J. Pan, A. C. Bovik, A. K. Moorthy, P. Panda, and K.-C. Yang, "In-capture mobile video distortions: A study of subjective behavior and objective algorithms," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 9, pp. 2061–2077, 2017.

[19] X. Yu, Z. Ying, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik, "Subjective and objective analysis of streamed gaming videos," *IEEE Trans. Games*, vol. 16, no. 2, pp. 445–458, 2024.

[20] H. Liu and I. Heynderickx, "A no-reference perceptual blockiness metric," in *IEEE Int. Conf. Acoust. Speech Signal Process.* IEEE, 2008, pp. 865–868.

[21] K. Rank, M. Lendl, and R. Unbehauen, "Estimation of image noise variance," *IEE Proc. Vis. Image Signal Process.*, vol. 146, no. 2, pp. 80–84, 1999.

[22] A. Ciancio, A. T. Da Costa, E. Da Silva, A. Said, R. Samadani, and P. Obrador, "Objective no-reference image blur metric based on local phase coherence," *Electron. Lett.*, vol. 45, no. 23, pp. 1162–1163, 2009.

[23] S. Ouni, E. Zagrouba, M. Chambah, and M. Herbin, "No-reference image semantic quality approach using neural network," in *IEEE Int. Symp. Signal Process. Inf. Technol.* IEEE, 2011, pp. 106–113.

[24] J. Zhang, T. M. Le, S. H. Ong, and T. Q. Nguyen, "No-reference image quality assessment using structural activity," *Signal Process.*, vol. 91, no. 11, pp. 2575–2588, 2011.

[25] D. Li, T. Jiang, and M. Jiang, "Unified quality assessment of in-the-wild videos with mixed datasets training," *Int. J. Comput. Vis.*, vol. 129, no. 4, pp. 1238–1257, 2021.

[26] J. Korhonen, Y. Su, and J. You, "Blind natural video quality prediction via statistical temporal features and deep spatial features," in *Proc. ACM Int. Conf. Multimedia*, 2020, pp. 3311–3319.

[27] Z. Zhang, W. Wu, W. Sun, D. Tu, W. Lu, X. Min, Y. Chen, and G. Zhai, "MD-VQA: Multi-dimensional quality assessment for ugc live videos," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 1746–1755.

[28] C. Feichtenhofer, H. Fan, J. Malik, and K. He, "Slowfast networks for video recognition," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 6202–6211.

[29] G. Denes, A. Jindal, A. Mikhailiuk, and R. K. Mantiuk, "A perceptual model of motion quality for rendering with adaptive refresh-rate and resolution," *ACM Trans. Graph.*, vol. 39, no. 4, pp. 133–1, 2020.

[30] R. Herzog, M. Čadík, T. O. Aydčin, K. I. Kim, K. Myszkowski, and H.-P. Seidel, "NoRM: No-reference image quality metric for realistic image synthesis," *Comput. Graph. Forum*, vol. 31, no. 2pt3, pp. 545–554, 2012.

[31] P. G. Barten, "Formula for the contrast sensitivity of the human eye," in *Image Qual. Syst. Perform.*, vol. 5294. SPIE, 2003, pp. 231–238.

[32] T. O. Aydin, M. Čadík, K. Myszkowski, and H.-P. Seidel, "Video quality assessment for computer graphics applications," *ACM Trans. Graph.*, vol. 29, no. 6, pp. 1–12, 2010.

[33] T. O. Aydin, R. Mantiuk, K. Myszkowski, and H.-P. Seidel, "Dynamic range independent image quality assessment," *ACM Trans. Graph.*, vol. 27, no. 3, pp. 1–10, 2008.

[34] K. Wolski, D. Giunchi, N. Ye, P. Didyk, K. Myszkowski, R. Mantiuk, H.-P. Seidel, A. Steed, and R. K. Mantiuk, "Dataset and metrics for predicting local visible differences," *ACM Trans. Graph.*, vol. 37, no. 5, pp. 1–14, 2018.

[35] L. Yang, S. Liu, and M. Salvi, "A survey of temporal antialiasing techniques," *Comput. Graph. Forum*, vol. 39, no. 2, pp. 607–621, 2020.

[36] "Epic games: Unreal engine 4," www.unrealengine.com, 2022.

[37] Epic Games. (2023) The most powerful real-time 3D creation tool - Unreal Engine. [Online]. Available: www.unrealengine.com/en-US

[38] AMD, "FidelityFX super resolution 2.0," gpuopen.com/fidelityfx-superresolution-2/, 2022.

[39] E. Liu, "DLSS 2.0-image reconstruction for real-time rendering with deep learning," in *GPU Technol. Conf.*, 2020.

[40] Epic, "Temporal super resolution," docs.unrealengine.com/5.2/en-US/temporal-super-resolution-in-unreal-engine/, 2022.

[41] Z. Tu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "UGC-VQA: Benchmarking blind video quality assessment for user generated content," *IEEE Trans. Image. Process.*, vol. 30, pp. 4449–4464, 2021.

[42] Y. Wang, S. Inguva, and B. Adsumilli, "YouTube UGC dataset for video compression research," in *IEEE 21st Int. Workshop Multimedia Signal Process.* IEEE, 2019, pp. 1–5.

[43] ITU-R, "Recommendation itu-r bt.500-14," www.itu.int/dms_pubrec/itu-r/rec/bt/R-REC-BT.500-14-201910-I!!PDF-E.pdf, 2020.

[44] R. K. Mantiuk, A. Tomaszewska, and R. Mantiuk, "Comparison of four subjective methods for image quality assessment," *Comput. Graph. Forum*, vol. 31, no. 8, pp. 2478–2491, 2012.

[45] L. T. DeCarlo, "On the meaning and use of kurtosis." *Psychol. Methods*, vol. 2, no. 3, p. 292, 1997.

[46] G. Le Moing, J. Ponce, and C. Schmid, "Dense optical tracking: Connecting the dots," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 19 187–19 197.

[47] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[48] W. Wen, M. Li, Y. Zhang, Y. Liao, J. Li, L. Zhang, and K. Ma, "Modular blind video quality assessment," *arXiv preprint arXiv:2402.19276*, 2024.

[49] L. Siqi and L. Yunzhen, "Introducing snapdragon game super resolution," www.qualcomm.com/news/onq/2023/04/introducing-snapdragon-game-super-resolution, 2023.

[50] Pixelworks, "Pixelworks X7 Gen 2 processor," www.pixelworks.com/media/products, 2024.

[51] AMD, "FidelityFX super resolution 1.0," gpuopen.com/fidelityfx-superresolution/, 2021.

[52] S. Yang, Y. Zhao, Y. Luo, H. Wang, H. Sun, C. Li, B. Cai, and X. Jin, "MNSS: Neural supersampling framework for real-time rendering on mobile devices," *IEEE Trans. Vis. Comput. Graph.*, vol. 30, no. 7, pp. 4271–4284, 2024.

[53] S. Yang, Q. Z. Zhu, J. Zhuge, Q. Qiu, C. Li, Y. Yan, H. Xu, L.-Q. Yan, and X. Jin, "Mob-FGSR: Frame generation and super resolution for mobile real-time rendering," in *Proc. ACM SIGGRAPH Papers*, 2024, pp. 1–11.

[54] Netflix, "VMAF - video multi-method assessment fusion," github.com/Netflix/vmaf, 2024.


**Sipeng Yang** received the BSc degree from Southeast University, P.R. China, in 2018. He is currently a PhD candidate in the State Key Lab of CAD&CG, Zhejiang University, China. His research interests include computer graphics and machine learning.

**Jiayu Ji** is currently pursuing a BSc degree in Computer Science and Technology at Zhejiang University, Hangzhou, P.R. China. His research focuses on deep learning, computer graphics, and computer vision.

**Qingchuan Zhu** received the BSc degree from Zhejiang University, P.R. China, in 2022. He is currently pursuing an MSc degree in Software Engineering at the State Key Lab of CAD&CG, Zhejiang University, China.

**Zhiyao Yang** received a BEng degree and D.Eng degree from Jilin University, P.R. China, in 2012 and 2017, respectively. He is a Senior Multimedia System Engineer in OPPO Guangdong Mobile Communications Co., Ltd. He is in charge with developing visual effect algorithms and functions, leading the end-side implementation of OPPO's super-resolution and visual enhancement features, and responsible for the application of HDR video technology. He also has been involved in a range of technical projects, these include intelligent video noise reduction and image inpainting, intelligent video anomaly detection, automated intelligent evaluation of video/game quality, and multi-view video applications.

**Xiaogang Jin** received a BSc degree, an MSc degree, and a PhD degree from Zhejiang University, P.R. China, in 1989, 1992, and 1995, respectively. He is a professor in the State Key Laboratory of CAD&CG, Zhejiang University. His current research interests include digital human animation, cloth animation and virtual try-on, digital portrait editing, AIGC and its applications, text2animation, neural supersampling for mobile devices, virutal reality, and computer games. He received an ACM Recognition of Service Award in 2015 and the Best Paper Awards from CASA 2017, CASA 2018 and GMP 2024. He is a member of the IEEE and the ACM.