Beyond Pixels: Employing PSNR、SSIM and VMAF for Comprehensive Video Quality Assessment

Abstract—**While PSNR (Peak Signal-to-Noise Ratio) is a straightforward and widely used metric, its lack of consideration for the spatial and temporal complexities of human visual perception limits its effectiveness in accurately gauging overall picture quality. In typical JPEG or MJPEG video codecs, PSNR measurements are done frame by frame. As a result, the PSNR closely mirrors the video content, rising significantly with easily compressible content and dropping with more detailed scenes. This creates a PSNR distribution that fluctuates quite frequently over time, often more so than the GOP (Group of Pictures) frequency seen in typical video codecs. Given this, how picture codecs, which are supposedly less sophisticated than advanced video codecs, manage to deliver stable visual experiences? In this short article, we emphasize it's essential to utilize more sophisticated measurements like SSIM (Structural Similarity Index) or VMAF (Video Multimethod Assessment Fusion) for a more holistic approach to video quality testing. These methods evaluate not just at the pixel level, but also take into account the surrounding pixels and their temporal relationships.**

Keywords—JPEG, MJPEG, SSIM, VMAF, PSNR

I. INTRODUCTION

The phenomenon of PSNR fluctuations in picture codecs, leading to unstable yet often imperceptible visual effects, has fueled skepticism about the reliability of PSNR as a metric for assessing video quality, highlighting the need for more comprehensive measures that can accurately reflect human visual perception. The journey through video quality assessment metrics has therefore evolved from PSNR, an objective metric evaluating pixel-level differences, to SSIM, which incorporates human visual perception factors such as texture and luminance, and ultimately to VMAF, a comprehensive model developed by Netflix in 2016, integrating machine learning to more closely align with human visual system's interpretation of video quality, marking a significant advancement in the pursuit of accurately measuring and ensuring high-quality video content in the digital age [1].

Is it accurate to say that the fluctuations in PSNR, similar to GOP PSNR drops in video content, don't necessarily indicate a decrease in perceived video quality? In other words, transitions from high to low PSNR (or vice versa) are perceived similarly by human eyes. Therefore, would it be incorrect to interpret the peak PSNR in picture codecs as a performance advantage? Could these fluctuations actually suggest instability or a less reliable performance when using picture codecs for video purposes? The human visual system doesn't always perceive changes in PSNR in the way we might expect based on the numbers alone. Fluctuations in PSNR, especially those resembling GOP PSNR drops, are not always indicative of a noticeable decline in video quality from the viewer's perspective. Consequently, using peak PSNR values to claim superiority of picture codecs in video applications can be misleading. These fluctuations might better be interpreted as a sign of the codec's instability or inconsistency in performance when used for video, rather than a definitive measure of enhanced performance. In other words, not only is PSNR insufficient as a measure for assessing video codec performance, but it also tends to provide misleading information when evaluating the quality of video codecs. This is precisely why measurements like SSIM and VMAF are more effective for evaluating video codec performance. Unlike PSNR, SSIM and VMAF take into account essential spatial and temporal information. For a high-quality video codec, we often observe a more stable or consistent trend in the SSIM or VMAF scores. This level of stability is not always evident in video codecs based on picture codecs, further highlighting the superiority of SSIM and VMAF as evaluation tools.

II. PICTURE CODEC AS A VIDEO CODEC AND ITS IMPLICATIONS

Frame-by-Frame Compression:
Picture codecs like JPEG compress each frame independently, without considering the temporal relationship between frames. This approach is fundamentally different from typical video codecs, which leverage similarities between successive frames to enhance compression efficiency.
Impact on PSNR:
PSNR measures the peak error between the original and compressed image. When using a picture codec for video, PSNR can fluctuate significantly from frame to frame. Scenes with less detail might compress well, showing high PSNR, while more complex scenes could result in a lower PSNR.
Effect on SSIM:
SSIM assesses the visual impact of changes in structural information, luminance, and contrast. Since picture codecs handle each frame in isolation, the structural integrity compared to the original can vary greatly across frames, leading to SSIM fluctuations.
Variation in VMAF:
VMAF, designed to reflect human perception, can also exhibit fluctuations when a picture codec is used for video. This is because VMAF takes into account factors like temporal pooling, which are impacted when frames are compressed without considering temporal relationships.

Why This Approach Leads to Fluctuations?
Lack of Temporal Coherence:
Without considering the temporal coherence between frames, each frame is a new challenge for the codec. This leads to inconsistency in quality across frames, as the codec does not

utilize information from adjacent frames to optimize compression.

Scene Complexity Variation:

In videos, the complexity of scenes can change rapidly. Picture codecs are not designed to adapt to these changes efficiently when used for video compression, resulting in varying quality metrics across frames, especially when the bandwidth is limited to be available.

Human Perception Sensitivity:

Viewers are generally more sensitive to fluctuations in video quality than to a consistently lower quality level. The inconsistency in quality metrics when using a picture codec for video is likely to be more perceptible and potentially distracting to viewers.

Using a picture codec like JPEG for video compression is inherently less efficient due to its frame-by-frame approach and lack of temporal awareness. This leads to notable fluctuations in key quality metrics like PSNR, SSIM, and VMAF, potentially degrading the viewer experience. These fluctuations highlight the importance of using codecs specifically designed for video to ensure consistent quality and efficient compression.

III. Measurement Formulas

A. SSIM (Structural Similarity Index) formulae

$$SSIM(x,y) = \frac{((2 * \mu x * \mu y + c1) * (2 * \sigma xy + c2))}{((\mu x^2 + \mu y^2 + c1) * (\sigma x^2 + \sigma y^2 + c2))}$$

Where:

x, y are the two windows of an image.

$\mu x, \mu y$ are the average of x, y.

$\sigma x^2, \sigma y^2$ are the variance of x, y.

σxy is the covariance of x, y.

$c1, c2$ are two variables to stabilize the division with a weak denominator.

B. PSNR (Peak Signal-to-Noise Ratio) formulae

$$PSNR = 20 * log10(MAX_I / \sqrt{MSE}))$$

Where:

MAX_I is the maximum possible pixel value of the image.

MSE is the Mean Squared Error between the original and compressed image.

PSNR's primary shortfall is its inability to adequately account for spatial and temporal factors that significantly affect perceived image quality.

C. VMAF (Video Multi-Method Assessment Fusion) formulae

VMAF is a full-reference video quality assessment method. While the exact formula involves complex machine learning models, a simplified representation involves integrating several metrics:

Feature Extraction: Features like Detail Loss Metric (DLM), Motion2, Visual Information Fidelity (VIF), and Additive Distortions Metric are extracted from both reference and distorted videos.

Feature Integration: These features are then combined using a *machine learning* model, often a Support Vector Machine, trained on a dataset rated by human viewers.

Quality Score Output: The model outputs a quality score, typically ranging from 0 to 100, where higher scores indicate better video quality.

The actual computation of VMAF in practice requires specific software, such as FFmpeg, to input reference and test video files and output the VMAF score. Due to its complexity and reliance on machine learning, the formulae is not as straightforward as traditional metrics like PSNR or SSIM [2-7]. In short, VMAF goes a step further by combining several metrics (which can include SSIM or metrics similar to SSIM) along with a machine learning model trained on video sequences rated by human viewers. This approach allows VMAF to consider a wider range of factors that affect perceived video quality, including temporal artifacts and other complex degradation types that SSIM might not fully capture.

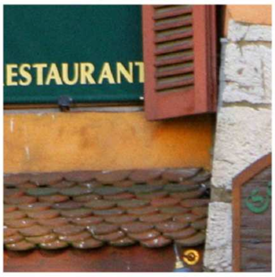
In practice, PSNR, SSIM and VMAF can be used complementarily. SSIM can provide quick and efficient assessments of *structural* similarities, while VMAF can be employed for a more thorough evaluation, especially in final stages of quality assurance or more detailed analysis.

IV. Examples without considering the accuracy of human perception

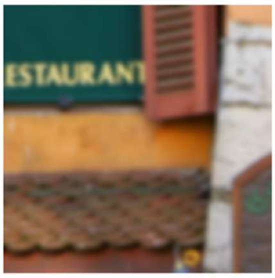
- PSNR

Both of the images below have an average PSNR score of 32 even though, subjectively, the image on the left is more distinguishable and there is much less detail visible in the blurry image on the right [1].

Noisy Original
PSNR: 32.21

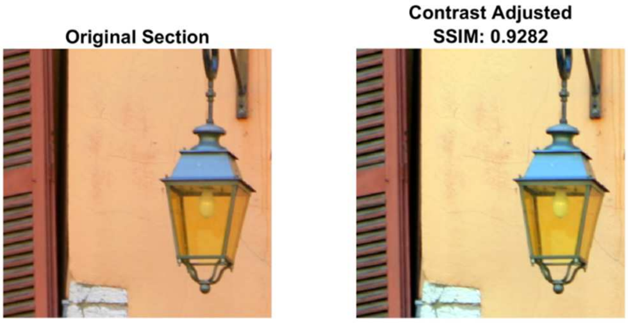


Blurred Image
PSNR: 32.16



- SSIM

SSIM is sensitive to any kind of structural changes, like the stretching of an image, rotations, or similar distortions. It is also affected by blockiness and blurriness. SSIM is also not the best at evaluating changes in image hue and similar factors. For example, the image on the left below is the original reference while the image on the right—which has completely different colors and is nowhere near the reference image—has an SSIM score of 0.93, which is still a very high SSIM score [1]. Please note that these types of distortions are not typical in high-quality video codecs with enough bitrate. SSIM remains a useful metric for evaluating the performance of high-quality video codecs



Numerous obstacles exist in utilizing full-reference metrics, notably regarding their accuracy. This is precisely why we discuss an array of these metrics. VMAF developed by Netflix, integrates various metrics, including SSIM, into a single framework to assess video quality, offering a more comprehensive analysis than traditional metrics. Unlike SSIM and PSNR, which are relatively simple calculations on still images, VMAF is designed for video and takes into account temporal factors (motion, frame rate, etc.) and spatial factors (resolution, detail, etc.). This complexity means creating VMAF paradoxes would require manipulating these factors in a video, not just a single frame or image. While VMAF provides deeper insights into video quality by considering factors closer to human visual perception, it's important to note that no metric is infallible. Although VMAF tends to be more reliable than PSNR, particularly in scenarios where PSNR may not correlate well with perceived quality, it's not entirely foolproof. The effectiveness of VMAF can vary depending on the content type, encoding settings, and the specific context in which it is used. Therefore, while VMAF is a significant advancement in video quality assessment, it should be used as part of a broader set of tools and considerations for evaluating video quality.

V. COMMON CONCERN ON QUALITY DIPS FROM VIDEO CODEC'S GOP

The concept of GOP (**Group of Pictures**) in video codecs is crucial for understanding video compression and quality. GOP refers to a collection of successive pictures within a coded video stream. A typical GOP starts with an I-frame (Intra-coded frame), followed by a series of P-frames (Predictive-coded frames) and B-frames (Bi-directionally predictive-coded frames).

Quality Dip in GOPs: It's true that a quality dip can occur at the beginning of a GOP. This is because I-frames are compressed without reference to other frames, which can lead to a higher level of compression artifacts compared to P and B frames, which use data from surrounding frames for more efficient compression.

Perceptual Impact: However, this dip in video quality might not always be perceptually significant. Human visual perception is quite complex, and factors such as the spatial and temporal masking effects can make these dips less noticeable. For instance, in scenes with high motion or complexity, viewers are less likely to perceive a decrease in quality [6].

Adaptive GOP Structures: There are ongoing advancements in adaptive GOP structures, where the GOP size and pattern are adjusted based on the video content, potentially minimizing the impact on perceived quality while maximizing compression efficiency [8].

GOP and Streaming: For video streaming, GOP structure plays a critical role in balancing the video quality and compression ratio. The choice of GOP length can affect the video stream's distortion sensitivity [9].

The perceptual impact of these dips, however, varies. Factors like the complexity of the video content, the viewer's attention, and the viewing environment can influence how noticeable these dips are. For instance, in high-motion or complex scenes, the quality dips might be less perceptible due to the viewer's focus being distributed across various elements [10]. Moreover, the recovery time from these dips can be attributed to the encoding efficiency of subsequent P and B frames. Since these frames are encoded using data from surrounding frames, they tend to restore quality more quickly, as they benefit from the temporal redundancy in the video sequence [8].

VI. COMPARING QUALITY METRIC FLUCTUATIONS IN JPEG-BASED AND ADVANCED VIDEO CODECS

The primary objective of this experiment is to methodically compare the fluctuations in key video quality metrics – PSNR 、SSIM and VMAF- between JPEG-based codecs and the video codec. These metrics are indispensable tools for objectively assessing video quality, offering insights into the visual fidelity and perceptual integrity of compressed video content. By examining how JPEG-based codecs and video codec perform, this experiment aims to shed light on the efficiency and effectiveness of these codecs in maintaining consistent video quality. Below is the setup for how to run this experiment.

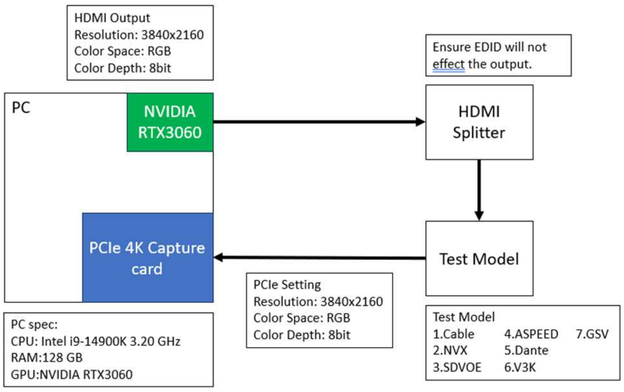


Fig 1. Experimental Setup

A. Selection of Video Sample:

In this section, we conduct a series of evaluations on a 10-second video at 4K/30 that varies in complexity, motion, and texture as shown below [11].

Fig 2. Snapshot of the input video with frame number label

B. *Codec Configuration*

For the experiment, the configurations of both the JPEG-based codec and the proprietary video codec were carefully set up to ensure an effective and fair comparison:

JPEG-Based Codec Configuration:

Bitrate Setting: The JPEG-based codec was configured with a high bitrate capacity, up to 800 Mbps. This setting is designed to test the codec's performance in a scenario where bandwidth is not a significant constraint.

Frame-by-Frame Compression: Unlike typical video codecs, JPEG-based codecs compress each frame independently as a single image, without considering temporal relationships between frames. This characteristic is crucial in our analysis, as it directly impacts the fluctuations in quality metrics.

Proprietary Codec Configuration:

Bitrate Setting: The proprietary video codec was set to operate at a maximum bitrate of up to 400 Mbps. This limit is to assess the codec's efficiency and effectiveness at a comparatively lower bitrate.

Adaptive Bitrate and GOP Structure: The proprietary video was configured to use adaptive bitrate streaming with a GOP size of 120 frames. This setup allows for the evaluation of how proprietary video codec handles changes in video scenes and its impact on the consistency of the quality metrics. The adaptive bitrate enables the codec to adjust the bitrate dynamically according to the complexity of the video content. This configuration sets the stage for examining how each codec performs under different bitrate settings and structural constraints. The distinct approaches of the JPEG codec (frame-by-frame compression) and the proprietary video codec (adaptive bitrate with specified GOP structure) will provide insights into how these factors influence the stability of PSNR, SSIM and VMAF in video compression scenarios.

C. *Data Presentation and Quality Metric Analysis*

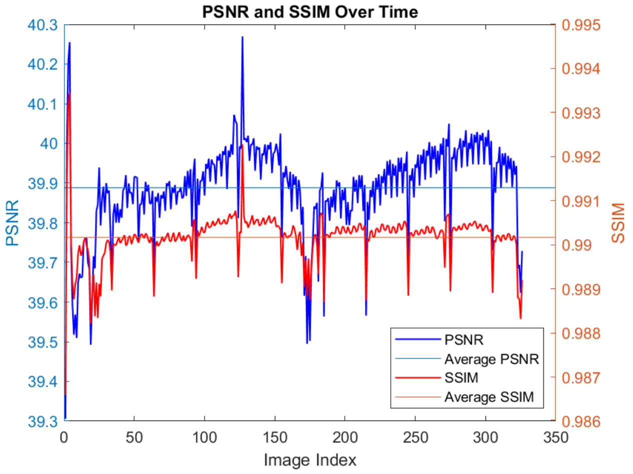


Fig 3. PSNR & SSIM Distribution of JPEG based Video Codec

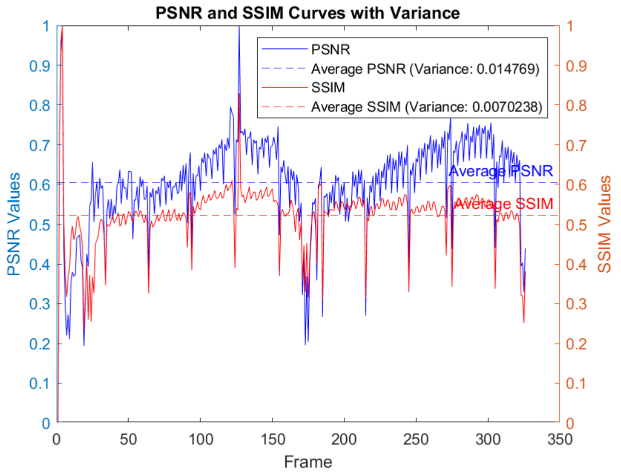


Fig 4. Normalized PSNR & SSIM Distribution of JPEG based Video Codec

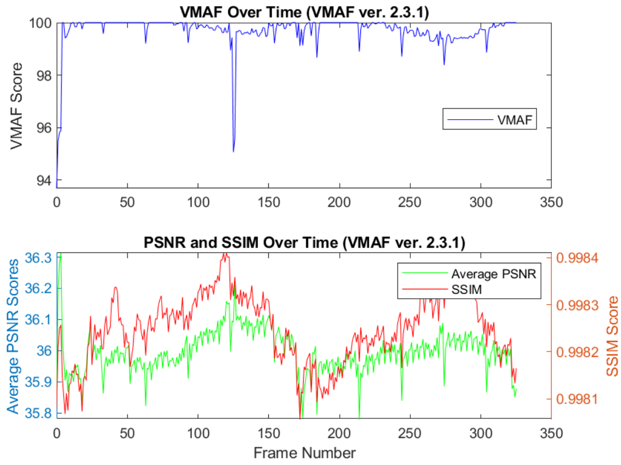


Fig 5. PSNR & SSIM & VMAF Distribution of JPEG based Video Codec

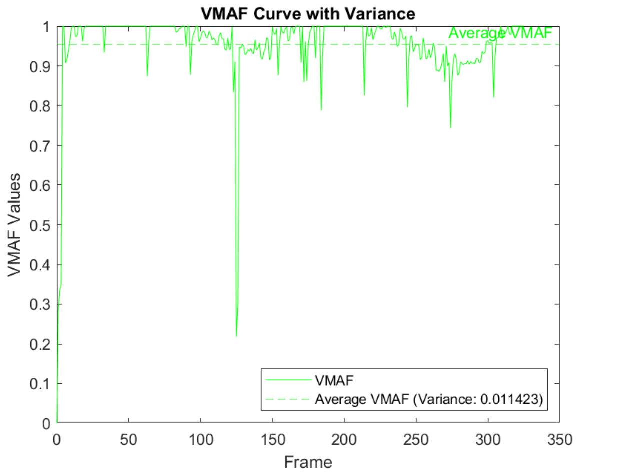


Fig 6. Normalized VMAF Distribution of JPEG based Video Codec

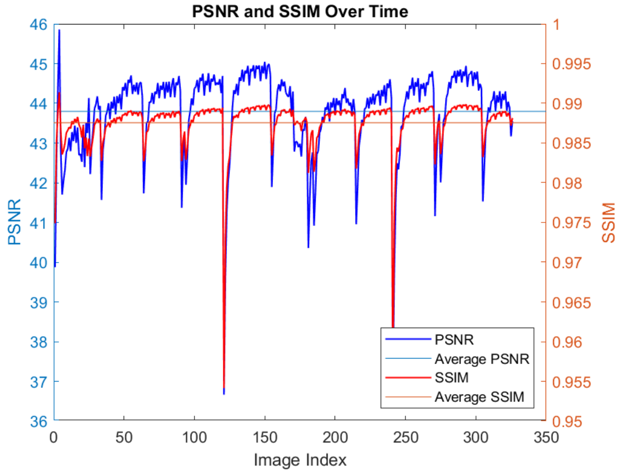


Fig7. PSNR & SSIM Distribution of the proprietary video codec

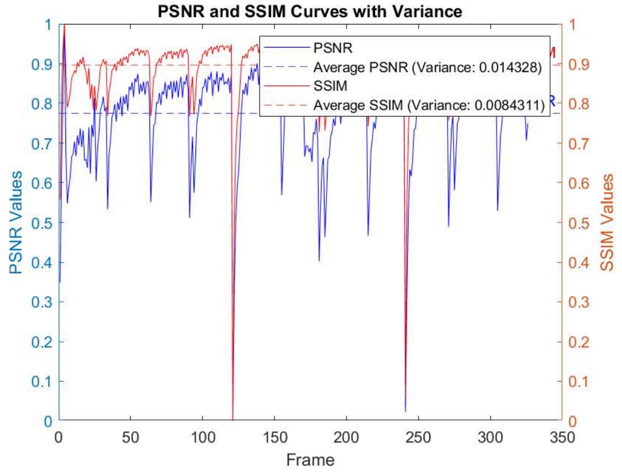


Fig 8. Normalized PSNR & SSIM Distribution of the proprietary video codec

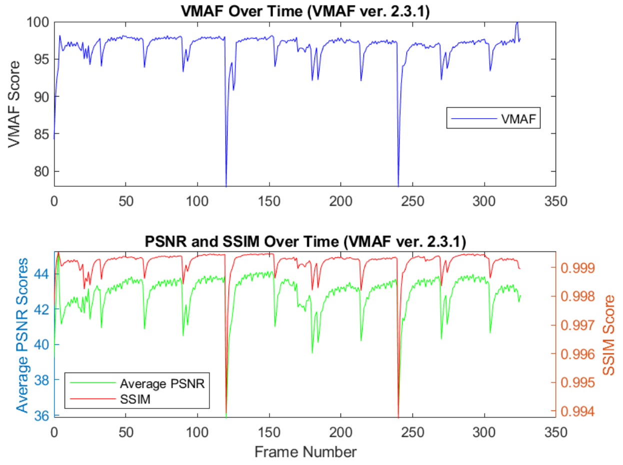


Fig 9. PSNR & SSIM & VMAF Distribution of JPEG based Video Codec

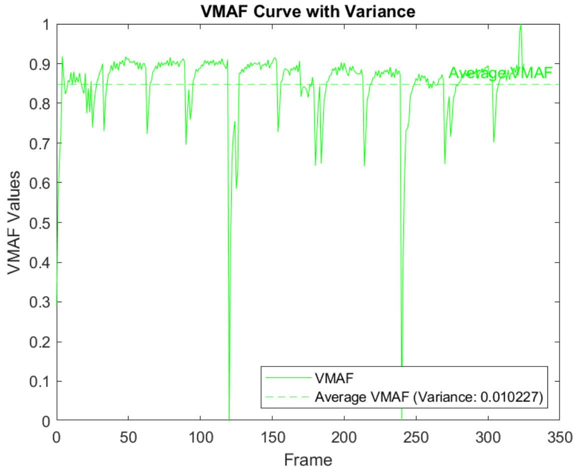


Fig 10. Normalized VMAF Distribution of the proprietary Video Codec

Fluctuations	JPEG Codec			Proprietary Video Codec		
	PSNR	*SSIM*	*VMAF*	*PSNR*	*SSIM*	*VMAF*
Nomalized Ratio	**1.47%**	**0.7%**	**1.14%**	**1.43%**	**0.84%**	**1.02 %**

Table I. Curve Fluctuations

In this experimental analysis, both JPEG and the proprietary video codec exhibit a similar level of perceptual instability, despite the proprietary video codec regularly encountering dips in Group of Pictures (GOP) measurements. However, overall, the proprietary video codec demonstrates greater stability across a broader range of metrics compared to JPEG codecs. This is primarily due to the proprietary video codec 's utilization of temporal information, which significantly contributes to its more stable output.

The patterns observed in VMAF and PSNR values across video frames indicate that JPEG codecs, lacking advanced temporal compression mechanisms like those found in P or B frames, are more susceptible to fluctuations influenced by the specific content of individual frames. The reliance on single-frame content leads to the performance of JPEG codecs being highly variable and dependent on the content itself. In contrast, the proprietary video codec's ability with existing GOP dips to leverage temporal data allows for a smoother transition and consistency between frames. This advantage is especially noticeable at lower bitrates, where maintaining consistent quality becomes challenging. As a result, the proprietary video codec is less prone to perceptual instability and are better suited for scenarios where maintaining a uniform video quality is essential.

It's important to note that this analysis serves as a comprehensive example to examine PSNR, SSIM, and VMAF metrics. While it provides valuable insights, it cannot cover all possible scenarios. Nonetheless, it offers a general idea about the performance capabilities of video and picture-based codecs, highlighting their strengths and limitations in various contexts. This analysis thus forms a useful reference point for understanding codec performance, though it should be considered as part of a broader assessment when evaluating video and image compression technologies.

VII. Evolving Video Quality Metrics in the Era of 4K and 8K: The Role and Impact of the Video Codec

In the evolving landscape of video streaming and compression, the choice of codec plays a pivotal role in determining the quality of the viewer experience. Recent discussions and observations have highlighted the superiority of the advanced video code, particularly when compared to older technologies like JPEG/MJPEG. This short article offers significant insights into the field of video quality assessment. Here are the key insights:

Relevance in High-Resolution Video Streaming: With the proprietary video codec supporting up to 4K or even 8K resolution, the paper's discussion on the effectiveness of advanced codecs becomes highly relevant. As consumers and industries move towards higher resolution content, understanding how codecs perform in terms of quality metrics (PSNR, SSIM, VMAF) is crucial.

Emphasis on Advanced Quality Metrics: The paper's focus on advanced quality assessment metrics like SSIM and VMAF gains additional importance. In high-resolution formats like 4K, the limitations of traditional metrics like PSNR become more pronounced, making the case for more sophisticated metrics even stronger.

Codec Efficiency and Performance: The paper highlights the efficiency of the proprietary video codec in maintaining stable quality metrics across frames. This insight is particularly valuable for ultra-high-definition videos, where efficient compression without sacrificing quality is paramount due to the immense data size.

Future Codec Development: The insights provided in the article can guide future developments in codec technology, especially in optimizing codecs for ultra-high-definition content. The evaluation of codecs in terms of SSIM and VMAF could become standard practice in codec development and assessment.

Practical Guidance for Industry Professionals: For professionals in the video production, streaming, and broadcasting industries, the article provides essential guidance on the choice of codecs for delivering high-quality 4K or 8K content. The comparative analysis with JPEG and other codecs becomes a valuable reference point.

Adoption and Standardization: The article's insights could influence the adoption and standardization of codecs in the industry, particularly for platforms and services that aim to provide high-definition content video.

A crucial aspect of video quality assessment is the stability of metrics like PSNR, SSIM, and VMAF. The proprietary video codec consistently shows fewer fluctuations in these scores, suggesting a more uniform quality of compression. This stability is not just a technical superiority but also translates to a better viewer experience. It's important to note that abrupt changes in these metrics, whether increases or decreases, can negatively impact viewer perception. In this regard, the proprietary video codec's performance is commendable, maintaining quality even in varied scene complexities. This article delves into the reasons behind this superiority, focusing on key aspects such as quality metric fluctuations and the handling of spatial and temporal information. the proprietary video codec stands out in the codec arena, especially when pitted against JPEG/MJPEG. Despite being able to operate at lower bitrates, the proprietary video codec demonstrates a remarkable ability to maintain consistent video quality. This is in contrast to JPEG/MJPEG, which, while being older and established codecs, show limitations in their compression techniques, particularly in dynamic scenes or complex textures. the proprietary video codec's ability to maintain stable quality metrics and handle spatial and temporal information efficiently makes itself a frontrunner in the quest for high-quality, efficient video streaming experiences.

VIII. References

[1] Audio Video Test Lab - Full-Reference Quality Metrics: VMAF, PSNR and SSIM

[2] github.com - vmaf/resource/doc/faq.md at master

[3] websites.fraunhofer.de - Calculating VMAF and PSNR with FFmpeg - Video-Dev

[4] medium.com - A practical guide for VMAF - Jina Jiayang Liu

[5] twitter.com - Introducing VMAF percentiles for video quality

[6] csdn.net - VMAF 原理学习笔记_视觉信息保真度

[7] stackoverflow.com - Right way to use vmaf with ffmpeg

[8] https://www.researchgate.net/figure/PSNR-vs-Bitrate-curves-comparing-fixed-GOP-sizes-to-the-proposed-adaptive-GOP-for_fig2_221126584

[9] https://repository.widyatama.ac.id/server/api/core/bitstreams/e9aeefb8-97ec-41f1-bd30-b91d22573666/content

[10] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8073500/

[11] https://drive.google.com/drive/u/0/folders/1KtRwxtwXMSu6Z4kmGeor3LZxSSqqdZkC?hl=zh-TW