“© 2026 IEEE. Personal use of this material is permitted. Permission from IEEE must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.”

TOWARD QUALITY ASSESSMENT OF 3D GAUSSIAN SPLATTING CODING

Joao Prazeres[†,‡], *Saeed Mahmoudpour*[◇,◇], *Stuart Perry*[⋆], *Manuela Pereira*[†,§], *António M.G. Pinheiro*[†,‡]

University of Beira Interior[†], Instituto de Telecomunicações[‡], NOVA LINCS[§], Vrije Universiteit Brussel[◇], IMEC[◇], University of Technology Sydney[⋆]

ABSTRACT

3D Gaussian Splatting (3DGS) recently emerged as an effective representation of radiance fields. These representations define 3D models of a scene based on a set of multiple views, resulting in large amounts of data, therefore, their practical application can benefit from efficient coding and representation. Performance evaluation of different coding solutions requires new quality evaluation protocols, based on subjective and objective models. While prior work on subjective quality assessment of Gaussian splatting primarily focused on evaluating rendering quality, the perceptual impact of compression-induced distortions remains largely unexplored. This paper presents a subjective and objective quality assessment of state-of-the-art 3DGS codecs, explicitly targeting coding artifacts rather than rendering effects. Six different scenes depicting diverse scenarios were rendered using Scaffold-GS as a common anchor. Two Scaffold-GS-based codecs, namely HAC++ and ContextGS, were evaluated at four distinct quality levels. A subjective quality assessment study was conducted to assess the perceptual degradation of compression. Additionally, the performance of a set of objective image quality metrics was analyzed.

***Index Terms*—** 3D Gaussian splatting, Coding, Subjective quality assessment, Objective quality metrics.

1. INTRODUCTION

3D Gaussian Splatting (3DGS) has rapidly evolved into a leading representation for radiance fields that provides real-time novel view synthesis [1], bridging the gap between classical point-based rendering and Neural Radiance Fields (NeRF) [2]. The 3D Gaussian Splatting framework proposed by Kerbl *et al.* [3] introduced an explicit, learnable set of anisotropic Gaussians optimized from multi-view imagery, providing a computationally more efficient alternative to the multiple NeRF solutions.

This work is funded by national funds through FCT – Fundação para a Ciência e a Tecnologia, I.P., and, when eligible, co-funded by EU funds under project/support UID/50008/2025 – Instituto de Telecomunicações, with DOI identifier: https://doi.org/10.54499/UID/50008/2025, and by the research projects MultImagePCQ: COMPETE2030-FEDER-00730300 and VITAL: CENTRO2030-FEDER-02584000

Despite the benefits, images created by Gaussian Splatting (GS) representations typically contain several visual artifacts, usually in the form of floating elements or geometric inconsistencies, thus negatively affecting the perceived quality of the rendered images.

This paper presents a framework developed by the JPEG Committee in its JPEG RF exploration activity for quality evaluation of compression artifacts in Gaussian Splatting methods by isolating codec effects under a shared Scaffold-GS [4] representation. It introduces a subjective quality assessment methodology considering two different trajectories for each scene. Furthermore, a benchmark of a set of objective quality metrics was performed based on the subjective results. Finally, a statistical analysis of the codec's behavior across bitrates and trajectories was performed.

For this study, a dataset of six diverse scenes was selected, rendered using Scaffold-GS [4] as the anchor representation, and encoded using two state-of-the-art Gaussian splatting codecs, namely HAC++ [5] and ContextGS [6], at multiple quality levels.

2. RELATED WORK

The scientific community has been active in researching quality assessment for radiance fields methods. Martin *et al.* conducted a subjective quality assessment study [7], using videos generated from several NeRF methods, complemented with an extensive objective study. Xing *et al.* [8] studied NeRF models that use structured representations such as voxel grids. It aimed to assess the effects of lossy compression on the perceived quality of rendered images, employing a subjective quality assessment based on the Double Stimulus Impairment Scale (DSIS). NeRF quality assessment was further explored by Liang *et al.* [9] by conducting a subjective quality assessment, employing the pairwise comparison methodology on NeRF synthesized images.

While NeRFs have been considered multiple times, Gaussian splatting methods have considerably fewer studies on subjective quality assessment. Yang *et al.* [10] proposed a graph-based Gaussian splatting compression method and defined its evaluation using the DSIS methodology. Zhang *et al.* [11] conducted a subjective quality assessment study between views generated by Gaussian splatting and by NeRF.



(a) *Bicycle* (b) *Bonsai* (c) *Garden* (d) *Playground* (e) *Train* (f) *Truck*

Fig. 1: Selected Dataset

Gaussian Splatting coding has also been a topic of interest, with several works being published. HAC++ [5] introduces hierarchical entropy models and hyperpriors for end-to-end compression of Gaussian parameters. ContextGS [6] improves coding efficiency by exploiting spatial dependencies between neighboring Gaussians through context-based probability modeling. Compact3DGS [12] focuses on practical storage reduction via pruning, low-rank factorization, and mixed-precision quantization, achieving large size savings with minimal quality loss. EAGLES [13] integrates quantization and bitrate penalties directly into GS optimization, enabling joint rate–distortion training. Finally, LightGaussian [14] and related lightweight variants reduce parameter counts through constrained Gaussian parameterization and feature sharing, targeting real-time and memory-limited deployment.

3. EXPERIMENT DESCRIPTION

3.1. Selected codecs

Hierarchical Attribute Coding++ (HAC++) [5] is a compression framework designed for 3D Gaussian Splatting (3DGS) that exploits the hierarchical structure of scaffold-based representations to improve coding efficiency. It organizes Gaussian parameters such as position, scale, rotation, opacity, and color across multiple resolution levels and applies progressive quantization and entropy coding to each hierarchy. By encoding coarse-level attributes first and refining them with finer residuals, HAC++ enables scalable rate–distortion control while preserving rendering fidelity.

ContextGS [6] is a context-aware entropy coding approach for 3DGS that models inter-parameter dependencies, achieving improved compression performance. It leverages spatial and attribute-level context, such as neighboring Gaussian properties and previously decoded parameters, to condition probability estimates in a learned entropy model. This approach allows more accurate symbol prediction and a lower bitrate while maintaining visual quality. By explicitly exploiting correlations among geometry, appearance, and opacity parameters, ContextGS achieves a superior rate–distortion performance, particularly at low bitrates, while remaining compatible with scaffold-based Gaussian representations.

Each codec was evaluated at four operating points by setting the rate–distortion parameter λ to $\{0.2, 0.06, 0.01, 5 \times 10^{-4}\}$ for HAC++ and $\{0.3, 0.12, 0.05, 10^{-3}\}$ for ContextGS. These settings define four different quality levels from low quality to very high quality.

These codecs were selected in the context of this work, as they provide for easy rate-distortion control. It is important to emphasize that this work does not intend to provide an evaluation of the existing coding solutions. Instead, this work targets the definition of a subjective quality evaluation model and an analysis of common image quality metrics.

3.2. Selected Reference

Scaffold-GS [4] is a structured extension of 3DGS that introduces a regularized scaffold of anchor points to guide the placement and optimization of Gaussian primitives. By constraining Gaussians to a hierarchical spatial backbone, it improves geometric stability, convergence behavior, and scalability to large scenes. Both selected codecs for this work are instantiation and coding models built upon the Scaffold-GS framework, encoding the same underlying Scaffold-GS representation. This shared structural foundation ensures consistent geometry and appearance across codecs, allowing observed performance differences to be allocated directly to the coding strategies rather than to disparities in representation learning and enabling fair comparisons.

3.3. Selected dataset

The selected data set consists of six different scenes depicting outdoor and indoor scenes, defined in the Mip-360 [15] and Tanks and Temples data set [16], shown in Fig. 1. The selected scenes are *Bicycle*, *Bonsai*, *Garden*, *Playground*, *Train* and *Truck*.

3.4. Visual Examples

Figure 2 shows visual examples of the *Bicycle* scene, coded with HAC++ [5] and ContextGS [6] at both the lowest and highest rates.

At the lowest rate, HAC++ shows strong degradation with dark speckle floaters, heavy blur in foliage and ground, and visible artifacts on the wheels, though the object remains recognizable. At the highest rate, HAC++ mainly exhibits residual blur in foliage and background, with slightly softened edges on rims and frame tubes.



(a) *Bicycle Scaffold-GS* (b) *Bicycle HAC++ R01* (c) *Bicycle ContextGS R01*

(d) *Bicycle HAC++ R04* (e) *Bicycle ContextGS R04*

Fig. 2: Example of coding artifacts for the selected codecs.

The lowest bitrate of ContextGS shows smoother foliage, fewer floaters, and more stable edges, though fine details like grass are still blurred. At the highest rate, ContextGS shows more spatially coherent noise, fewer floaters, and better edge preservation, with only mild blur in vegetation, yielding results closer to the reference.

3.5. Trajectory Selection

Typically, neural scene methods such as NeRF and 3DGS provide continuous camera trajectories to generate fly-through videos. Because Structure from Motion (SfM) [17] reconstructions (e.g., COLMAP) produce uneven camera coverage, with dense clusters and large under-sampled regions, test trajectories must be carefully designed to reveal parallax and stress novel viewpoints. Fully automatic path generation is unreliable, as it is not content-aware and often will create frames that pass through objects or focus on artifact-heavy regions.

For each scene, a set of key views was manually selected. Camera positions were interpolated linearly to ensure constant-velocity motion, while rotations were interpolated using SLERP [18] to guarantee smooth angular transitions. Two 10-second trajectories were generated per scene, one emphasizing well-sampled regions and one targeting under-sampled areas to expose reconstruction weaknesses. The selection of two different trajectories is important to understand how dependent is the quality variation from the selected trajectory. Even when focusing on under-sampled areas, effort was taken to ensure that the image did not contain a large number of artifacts induced by the Scaffold-GS rendering mechanism. Avoiding these rendering artifacts was very important to ensure that subjects concentrated their quality evaluation on coding artifacts only.

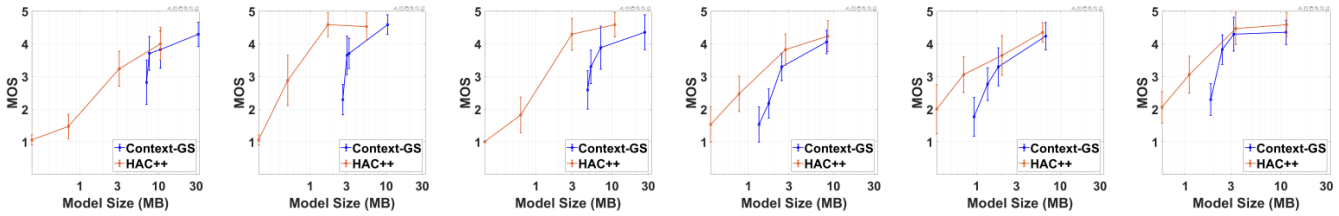
4. SUBJECTIVE QUALITY ASSESSMENT METHODOLOGY

The Double Stimulus Impairment Scale (DSIS) subjective protocol was employed. Subjects were first shown the anchor content (Scaffold-GS) 10-second sequence, followed by the coded 10-second sequence. They were then asked to rate the coded content using a five-level annoyance scale ranging from very annoying (1) to imperceptible (5), as defined in ITU-R BT.500-15 [19]. Prior to the evaluation, a training session was conducted for each subject. Six videos, covering distinct quality levels of a scene not included in the test, were shown to each subject, allowing them to become familiar with the coding artifacts and the evaluation procedure. Stimulus order was randomized to avoid consecutive presentation of coded versions of identical content. All tests were conducted in the Image and Video Technology Group's subjective laboratory at Universidade da Beira Interior, using a 10-bit 4K Eizo CG318 display under ITU-R BT.500-15 conditions [19]. Seventeen volunteer subjects (11 male, 6 female, aged 19–36) participated in the study. All the subjects were tested to ensure normal or corrected-to-normal vision using the Snellen[1] visual acuity test and absence of color blindness using the Ishihara[2] test.

The mean opinion score (MOS) results, together with the confidence intervals (CIs) for the subjective assessment, are

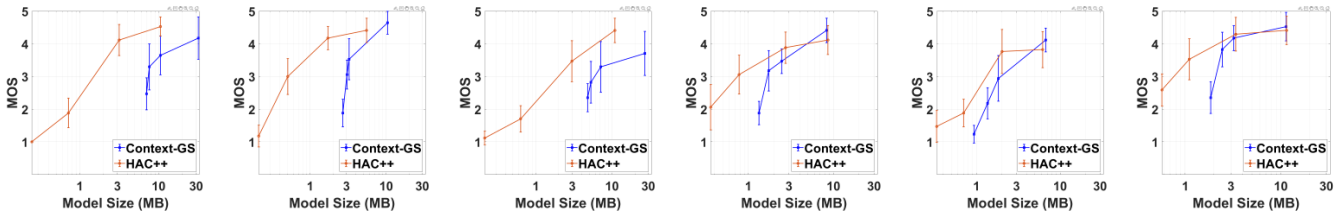
[1]https://visionscreening.zeiss.com/en-INT

[2]https://www.blindnesstest.com/ishihara-test/



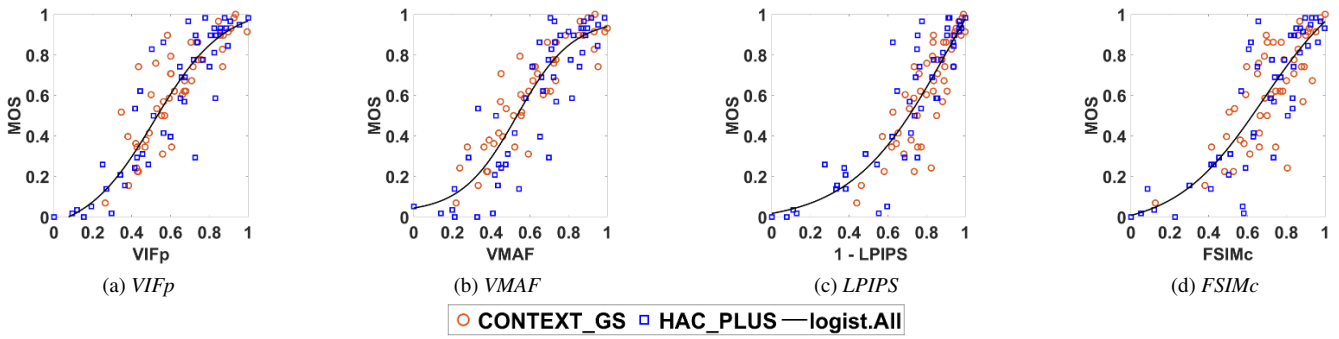
(a) *Bicycle* (b) *Bonsai* (c) *Garden* (d) *Playground* (e) *Train* (f) *Truck*

Fig. 3: MOS vs. Model Size with 95% CIs, considering a Gaussian distribution for trajectory 1.



(a) *Bicycle* (b) *Bonsai* (c) *Garden* (d) *Playground* (e) *Train* (f) *Truck*

Fig. 4: MOS vs. Model Size with 95% CIs, considering a Gaussian distribution for trajectory 2.



(a) *VIFp* (b) *VMAF* (c) *LPIPS* (d) *FSIMc*

Fig. 5: Logistic fitting for each metric.

reported for both trajectories in Fig. 3 and 4.

5. OBJECTIVE QUALITY ASSESSMENT

A set of objective quality metrics was selected, namely PSNR-Y, PSNR-YUV, SSIM [20], MS-SSIM [21], FSIMc [22], LPIPS [23], VIFp [24], VMAF [25], and IW-SSIM [26]. The considered image metrics were computed for each individual frame of the tested videos, using ScaffoldGS as a reference. The final metric value was the average of the individual values for all the frames of each video.

Furthermore, in order to assess the performance of the selected objective metrics, the statistical measures present in the ITU-T P.1401 recommendation [27], namely the Pearson Correlation Coefficient (PCC), the Spearman Ranked Order Correlation Coefficient (SROCC), the Root Mean Squared Error (RMSE), and the Outlier Ratio (OR) were computed. The quality scores predicted for each of the objective metrics were computed by applying a logistic fit function to the objective scores [27], as it is commonly done when benchmarking objective metrics [28].

6. DISCUSSION OF THE RESULTS

Figures 3 and 4 show the subjective quality results for each trajectory. Trajectory 1 focuses on well-sampled regions, while trajectory 2 focuses on under-sampled regions. The results show that both HAC++ and ContextGS achieve substantial coding gains while maintaining an acceptable perceptual quality. Complex scenes, containing fine geometry and foliage such as *Garden* and *Bonsai*, reveal a slower perceptual quality increase, while simpler scenes such as *Truck*

Table 1: Metric Performance

Metric Name	PCC	SROCC	RMSE	OR
PSNR-Y	0.640	0.608	0.229	0.781
PSNR-YUV	0.663	0.636	0.223	0.844
SSIM	0.729	0.707	0.204	0.792
MS-SSIM	0.790	0.775	0.183	0.688
LPIPS	0.873	0.862	0.144	**0.542**
IW-SSIM	0.822	0.806	0.169	0.698
FSIMc	0.828	0.805	0.166	0.646
VIFp	**0.887**	0.877	**0.138**	**0.542**
VMAF	0.886	**0.880**	**0.138**	0.656

and *Train* saturate at higher bitrates. It should be noted that the only non-monotonical behavior where a small MOS drop occurs is for trajectory 1 of the *Bonsai* scene coded with HAC++ at the R04 rate. The observation of R03 and R04 rate video sequences did not reveal any specific reason for this behavior. The quality is revealed to be very high in both cases.

The plots also reveal that trajectories that emphasize under-sampled regions yield systematically lower results for both codecs, confirming that artifacts are amplified in poorly sampled regions. Although the rendering artifacts are almost unperceived, the compression mechanism tends to augment them, resulting in lower perceived quality. These results highlight the importance of trajectory selection by considering under-sampled regions to enable a fair comparison of different coding solutions.

Table 1 shows the performance of the quality metrics when compared with the subjective quality scores. VIFp and VMAF achieve the best correlation scores (PCC/SROCC of 0.887/0.877 and 0.886/0.880, respectively). Overall, most metrics achieved a poor performance on predicting the subjective scores. It is important to emphasize that both reference and distorted images from the selected trajectories are rendered from a Gaussian splatting. This process is likely to create spatial drift or opacity instability that is poorly captured by traditional image quality metrics, as images under comparison might not be perfectly aligned.

Fig. 5 shows the logistic fitting results for the four best-performing metrics. Several operating points with similar objective scores differ meaningfully from the scores, especially at low bitrates, indicating that the usage of the typical metrics (PSNR, SSIM, and LPIPS) to optimize Gaussian splatting coding methods can lead to non-optimal perceptual outcomes.

The results of this study are publicly available[3].

6.1. Statistical Analysis

The reliability and discriminability of the subjective quality scores were statistically validated following standard practices for MOS-based subjective experiments [29]. The associated uncertainty was quantified using 95% confidence intervals (CI), and overall reliability was defined by the normalized confidence interval [30], obtained by dividing the mean CI by the full scale range.

[3]https://github.com/JpcPrazeres/GSC-AIPfQA

Table 2: Statistical validation of the subjective quality evaluation.

Statistical Criteria	**Value**
Normalized confidence interval (CI_{norm})	0.113
Effect size (Δ)	4.89
ICC	0.59
ICC 95% CI	[0.51, 0.67]

Inter-rater reliability was assessed using the Intraclass Correlation Coefficient (ICC) [31], corresponding to a two-way random-effects model with absolute agreement. Discriminability was evaluated using the effect size Δ [32], computed as the mean standardized difference between MOS values across all stimulus pairs, with the standard deviation estimated via bootstrap resampling (500 iterations).

As shown in Table 2, the normalized confidence interval of 0.113 indicates good reliability, while ICC reflects moderate inter-rater agreement. The large effect size (Δ = 4.89) reveals strong discriminability among the tested conditions.

7. CONCLUSIONS

This work presents a subjective quality evaluation framework for Gaussian Splatting coding distortions. The subjective experiment was designed to make the coding artifacts dominant against eventual rendering artifacts caused by the reconstruction mechanism of the Scaffold-GS anchor. Statistical analysis confirms good reliability and strong discriminability of the subjective data. Overall, the study establishes a reliable subjective assessment protocol for quality evaluation of Gaussian Splatting compression. The weak correlation between subjective scores and conventional objective metrics highlights the limitations of pixel-based measures for Gaussian Splatting and motivates subjective-driven evaluation approaches. The results show that the performance of both codecs is highly influenced by the sampling level of the original content. Under-sampled areas consistently produce lower perceived quality, although the rendering distortions without coding were kept at very low levels. Moreover, objective results show fair correlations of some metrics with the subjective ratings, while also suggesting the need for new metrics tailored to 3DGS compression artifacts.

8. REFERENCES

[1] Ben Fei et al., "3D Gaussian splatting as new era: A survey," *IEEE Trans. Vis. Comput. Graph.*, 2024.

[2] Ben Mildenhall et al., "NeRF: Representing scenes as

neural radiance fields," *Commun. ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[3] Bernhard Kerbl et al., "3D Gaussian splatting for real-time radiance field rendering," *ACM Trans. Graph.*, vol. 42, no. 4, pp. 139, 2023.

[4] Tao Lu et al., "Scaffold-GS: Structured 3D gaussians for view-adaptive rendering," in *Proc. CVPR*, 2024, pp. 20654–20664.

[5] Yihang Chen et al., "HAC++: Towards 100x compression of 3D gaussian splatting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 11, pp. 10210–10226, 2025.

[6] Yufei Wang et al., "ContextGS: Compact 3D gaussian splatting with anchor-level context model," *Adv. Neural Inf. Process. Syst.*, vol. 37, pp. 51532–51551, 2024.

[7] Pedro Martin et al., "NeRF view synthesis: Subjective quality assessment and objective metrics evaluation," *IEEE Access*, 2024.

[8] Yuke Xing et al., "Explicit-NeRF-QA: A quality assessment database for explicit NeRF model compression," in *Proc. VCIP*, 2024, pp. 1–5.

[9] Hanxue Liang et al., "Perceptual quality assessment of NeRF and neural view synthesis methods," *Comput. Graph. Forum*, vol. 43, no. 2, 2024.

[10] Qi Yang et al., "A benchmark for gaussian splatting compression and quality assessment," in *Proc. MMAsia*, 2024, pp. 1–8.

[11] Yuhang Zhang et al., "Evaluating human perception of novel view synthesis," *arXiv:2501.08072*, 2025.

[12] Joo Chan Lee et al., "Compact 3D gaussian representation for radiance field," in *Proc. CVPR*, 2024, pp. 21719–21728.

[13] Sharath Girish et al., "EAGLES: Efficient accelerated 3D gaussians with lightweight encodings," in *Proc. ECCV*, 2024, pp. 54–71.

[14] Zhiwen Fan et al., "LightGaussian: Unbounded 3D gaussian compression with 15x reduction and 200+ fps," in *Proc. NeurIPS*, 2024.

[15] Jonathan T. Barron et al., "Mip-NeRF 360: Unbounded anti-aliased neural radiance fields," *CVPR*, 2022.

[16] Arno Knapitsch et al., "Tanks and temples: Benchmarking large-scale scene reconstruction," *ACM Transactions on Graphics*, vol. 36, no. 4, 2017.

[17] Johannes L. Schönberger et al., "Structure-from-motion revisited," in *Proc. CVPR*, 2016, pp. 4104–4113.

[18] Young Kyun Jang et al., "Spherical linear interpolation and text-anchoring for zero-shot composed image retrieval," in *Proc. ECCV*, 2024, pp. 239–254.

[19] ITU-R BT.500, "Methodology for the subjective assessment of the quality of television pictures," 2023.

[20] Zhou Wang et al., "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, 2004.

[21] Zhou Wang et al., "Multiscale structural similarity for image quality assessment," in *Proc. Asilomar Conf. Signals Syst. Comput.*, 2003, pp. 1398–1402.

[22] Lin Zhang et al., "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, 2011.

[23] Richard Zhang et al., "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. CVPR*, 2018, pp. 586–595.

[24] H. R. Sheikh et al., "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, 2006.

[25] Zhi Li et al., "Toward a practical perceptual video quality metric," *Netflix Technology Blog*, 2016.

[26] Zhou Wang et al., "Information content weighting for perceptual image quality assessment," *IEEE Trans. Image Process*, vol. 20, no. 5, pp. 1185–1198, 2010.

[27] ITU-T P.1401, "Methods for statistical evaluation of objective quality prediction models," 2020.

[28] Philippe Hanhart et al., "Benchmarking of objective quality metrics for hdr image quality assessment," *EURASIP J. Image Video Process.*, 2015.

[29] Saeed Mahmoudpour et al., "Benchmarking subjective quality assessment methodologies for light field compression," *IEEE Transactions on Broadcasting*, 2026.

[30] Toshiko Tominaga et al., "Performance comparisons of subjective quality assessment methods for mobile video," in *Proc. QoMEX*. IEEE, 2010, pp. 82–87.

[31] David Liljequist et al., "Intraclass correlation–a discussion and demonstration of basic features," *PloS one*, vol. 14, no. 7, pp. e0219854, 2019.

[32] Herbert Aron David, *The Method of Paired Comparisons*, vol. 12, London, 1963.