

Beyond the Score: A Zero-Shot Vision–Language Judge for Interpretable, Artifact-Aware Quality Assessment of Traditional and Neural Video Codecs

Breno José Ribeiro de Vasconcelos

bjrv@cin.ufpe.br

Centro de Informática (CIn), Universidade Federal de
Pernambuco (UFPE)
Recife, PE, Brazil

Carlos André Guimarães Ferraz

cagf@cin.ufpe.br

Centro de Informática (CIn), Universidade Federal de
Pernambuco (UFPE)
Recife, PE, Brazil

ABSTRACT

Standard video-quality metrics quantify *how much* a compressed frame degrades — not *why*. Blocking, texture blur, ringing: the artifact type dictates the remedy, but every metric collapses them into a single number and remains silent about the cause. The problem deepens with neural codecs, which fail differently from traditional ones — smearing fine texture rather than introducing blocking — but no metric distinguishes the two. We asked whether off-the-shelf vision–language models, with no quality-specific training, can act as perceptual judges that *name* the artifact and *justify* their verdict. Querying **Claude Opus 4.8**, **GPT-5.5**, and **Gemini 3.5 Flash** zero-shot on 216 compressed sequences (traditional AV1/VVC and neural DCVC-FM/RT), we found they **outperform every classical frame-level metric** — PSNR (0.63), SSIM (0.59), LPIPS (0.62) — tracking human opinion at SROCC 0.71–0.74 on identical per-frame inputs, *without any training*. Beyond the score, they name the artifact: every blocking label was assigned to a traditional codec; not one was wrongly attributed to a neural one. Three independent AI vendors agreed on this, with inter-model correlation above 0.88. The one metric they do not surpass is VMAF — which, unlike the judges, integrates the *entire* video sequence. That gap is real, and we report it explicitly. Within the per-frame regime, however, these zero-shot judges surpass every classical frame-level metric and return something none of them can: a *reason* for the score.

KEYWORDS

video quality assessment, neural video codecs, vision–language models, LLM-as-judge, compression artifacts, perceptual metrics, VMAF

1 INTRODUCTION

Learning-based (neural) video codecs have closed, and on several operating points surpassed [11, 12, 19], the rate–distortion performance of standardized codecs. On the benchmark of Tavares et al. [19], the neural codec DCVC-FM saves up to 37.82% bitrate over HEVC (High Efficiency Video Coding) under a traditional BD-rate (Bjontegaard Delta rate) measurement (and 45.55% under a neural-codec measurement), and at a fixed bitrate improves quality by +1.86 dB PSNR (Peak Signal-to-Noise Ratio) and +10.3 VMAF

(Video Multi-Method Assessment Fusion). That same study shows the two families fail *differently*: traditional HEVC/VVC (Versatile Video Coding) [2] codecs introduce blocking and ringing, while neural codecs (DCVC-FM, DCVC-RT) preserve structure but *blur fine texture*. The *type* of degradation, not just its magnitude, now varies with the codec family.

These metrics, however, collapse quality into a single scalar. PSNR and SSIM (Structural Similarity Index) are signal-fidelity measures; VMAF [13] and LPIPS (Learned Perceptual Image Patch Similarity) [25] track perception better but are still *agnostic to the kind of artifact*: they report a number, not whether a frame was blocked, blurred, or color-shifted, nor *where*. More critically, fidelity-oriented metrics may misrank neural reconstructions whose texture blur is perceptually mild but numerically large [19]. Compounding this, codec comparisons (including [19]) often draw perceptual conclusions from objective metrics *without* validating them against human opinion, and no objective metric explains its own verdict.

When two codecs score the same VMAF but one blocks and the other blurs, an engineer choosing a deployment, debugging a rate-control decision, or triaging a quality regression cannot act on the scalar alone: the *kind* of failure dictates the fix (deblocking strength, texture synthesis, bit allocation), but it is precisely what the metric hides. Two developments, however, make a different approach practical. First, codec evaluation has itself become family-dependent (Sec. 2), so a distortion-type-aware tool is newly valuable. Second, using a strong model as an automatic “judge” has become a reliable, widely studied protocol in natural-language evaluation [26], and frontier vision–language models have demonstrated sensitivity to low-level visual detail [22], raising the question of whether they can serve as *interpretable* perceptual judges without any task-specific training.

We ask whether a multimodal vision–language model (VLM), used *off the shelf* and *without any task-specific training*, can serve as a perceptual judge that is *interpretable*: given a reference and a reconstructed crop, it emits a quality score, names the dominant artifact from a fixed taxonomy, and justifies the verdict in one sentence. We study three frontier commercial VLMs (Claude Opus 4.8, GPT-5.5, and Gemini 3.5 Flash) as zero-shot judges and validate every output against human MOS (Mean Opinion Score) from a public dataset covering both traditional and neural codecs, on *all* 216 sequences with no example selection. The specific findings and their scope are summarized in the contributions below; we report where the method works and, explicitly, where it fails.

Research questions.

- **RQ1.** How closely does a zero-shot VLM judge track MOS, relative to PSNR/SSIM/LPIPS and VMAF, on both traditional and neural codecs?
- **RQ2.** Does the judge’s dominant-artifact label discriminate codec families (blocking vs. texture-blur), a distinction no objective metric provides?
- **RQ3.** What does the judge add beyond the score (interpretability, cross-vendor agreement), and where does it systematically fail?

Contributions.

- (1) A **training-free, full-reference VLM judge** that, beyond a quality score, emits an *interpretable categorical artifact label* and a one-sentence justification — a capability absent from PSNR/SSIM/VMAF/LPIPS and from *trained* neural video quality assessment (VQA) models (e.g., CAMP-VQA, LMM-VQA). The differentiator is *interpretability and zero task-specific training*.
- (2) A **full, unbiased evaluation** on all 216 PVS (Processed Video Sequences) of AVT-VQDB-UHD-1-NVC, covering neural (DCVC-FM/RT) and traditional (AV1/VVC) codecs (*no example selection*), across three judges (Claude Opus 4.8, GPT-5.5, Gemini 3.5 Flash). On matched frame crops the judges significantly *outperform* PSNR, SSIM, and LPIPS (Williams’s test, $p < 0.01$) and trail only full-video VMAF; the blocking label is flagged exclusively for traditional codecs — a family-discriminative signal absent from any scalar metric.
- (3) An **open evaluation suite**: the **verbatim prompt text** (system prompt, score pass, and artifact pass), JSON schema, scoring scripts, and per-PVS results, released at <https://github.com/breno-sh/vlm-vqa-judge-artifact>, together with the methodological choices that make them replicable — native-resolution crops, ITU-R BT.500-anchored scoring, VQEG/ITU-T P.1401 nonlinear fit before PLCC (Pearson linear correlation coefficient), Shapiro–Wilk normality checks, per-content bootstrap CIs (SROCC) and Fisher-z CIs (PLCC), Williams’s test for dependent correlations, and paired *t*/Wilcoxon error tests.

2 RELATED WORK

2.1 Objective full-reference quality metrics

Full-reference (FR) quality assessment spans a spectrum of modeling assumptions. Signal-fidelity measures (PSNR and the Structural Similarity index SSIM [21]) are simple, fast, and routinely reported in codec comparisons, but they correlate only loosely with perception, especially for distortions that preserve structure while altering texture. Perceptual metrics close part of this gap: LPIPS [25] measures distance in the feature space of a pre-trained network, as validated against human similarity judgments, while DISTS (Deep Image Structure and Texture Similarity) [4] explicitly unifies *structure* and *texture* similarity and is, by design, tolerant of texture resampling — a property directly relevant to neural codecs that re-synthesize rather than copy texture (we discuss DISTS here for context but do not use it as a baseline, as it is not provided by the dataset for all PVS). VMAF [13] fuses elementary features

through a model trained on subjective data and is widely used for streaming-quality monitoring; its reproducibility has been independently studied [17]. Subjective ground truth itself is governed by ITU standards: BT.500 [8] and P.910 [10] for test methodology, and P.1401 [9] for the statistical evaluation of objective predictors (the logistic mapping, RMSE (root mean square error), and correlation analysis we adopt). Every one of these metrics emits a single scalar and is *agnostic to the type of distortion*: none can state whether quality dropped because of blocking, ringing, or texture blur, nor where in the frame — exactly the information a codec engineer needs when, as below, the failure mode now depends on the codec family.

2.2 Neural video codecs and their artifacts

Conditional-coding neural codecs such as DCVC-FM [12] and the real-time DCVC-RT [11] have matched or surpassed standardized codecs (AV1, VVC) on rate–distortion, and the comparative study of Tavares et al. [19] documents that the two families *fail differently*: traditional transform coding produces blocking and ringing, whereas neural codecs preserve global structure but blur or synthesize non-reference fine texture. This shift makes *distortion-type-aware* evaluation genuinely useful: a scalar calibrated on blocking-dominated content may misjudge texture-blur-dominated neural output. Several recent datasets address this gap: LEHA-CVQAD [6] and the AIM-2024 challenge [18] provide large compressed-VQA collections, and AVT-VQDB-UHD-1-NVC [15, 16] adds per-PVS MOS for neural and traditional codecs at UHD (ultra-high-definition), which is what enables the family-contrastive validation we perform. Recognizing these artifact differences, several trained neural and multimodal models have recently targeted compressed video quality directly.

2.3 Trained neural / LMM-based VQA

An active line of research adapts large multimodal models (LMMs) to quality assessment. Q-Bench [22] benchmarks the low-level visual perception of LMMs, and Q-Align [23] shows that teaching an LMM to emit *discrete text-defined levels* (rather than raw numbers) markedly improves scoring — a finding we operationalize in our ITU-anchored prompt. Q-Align is, however, a *fine-tuned* model (task-specific training on quality data), operates in a no-reference regime, and does not output artifact labels; it is therefore not a zero-shot baseline for our setting, though it represents a natural upper-bound comparison for future fine-tuned extensions of the approach. LMM-VQA [5] and VideoScore [7] train multimodal models for no-reference video and generated-video quality. CAMP-VQA [20] is similarly fine-tuned and, notably, attributes part of its difficulty with compressed content to “the absence of fine-grained perceptual annotations of artifact types” — the very capability our judge provides without additional annotation. Some trained multimodal systems likewise pair a quality score with a textual description (e.g., VideoScore [7], LMM-VQA [5]), but as task-specific rather than zero-shot commercial APIs, typically without a discrete artifact label and without an explicit traditional-vs-neural comparison that includes DCVC codecs. All require *task-specific training*; none provides a training-free, API-only judge that also classifies artifacts and contrasts traditional against neural codecs.

An alternative path is to skip fine-tuning entirely and use a strong frontier model as an evaluator directly: the LMM-as-judge paradigm now well established in natural-language evaluation.

2.4 LMM-as-judge and absolute vs. relative scoring

Using a strong model as an evaluator (“LLM/LMM-as-judge”) is now standard in NLP (Natural Language Processing) [26] and has been extended to visual quality assessment [22]. A consistent and well-documented finding is that such judges *rank* candidates far more reliably than they assign calibrated *absolute* scores, and that they exhibit biases (position, verbosity) that pairwise protocols mitigate [24, 26]. Wu et al. [24] systematically benchmark GPT-4V, Gemini, and other frontier models on IQA (Image Quality Assessment) tasks, finding that GPT-4V tracks human perception reasonably well but struggles with fine-grained discrimination — a limitation our ITU-anchored prompt and multi-model ensemble are designed to address. Accordingly, rank correlation (SROCC/KROCC) is our primary evidence; PLCC is reported only after the ITU-T P.1401 logistic mapping. Two-alternative forced-choice (2AFC) prompting [27] can further mitigate absolute-score calibration issues; we adopt absolute scoring here (216 PVS $\times$ 3 judges makes pairwise comparison impractical at scale) and note a Bradley–Terry formulation as future work.

Research gap. To our knowledge, no prior work combines all of: *zero-shot* use of *commercial* VLMs through an API (no training), *full-reference* operation, an explicit *traditional-vs-neural* comparison including DCVC-FM/RT, a *categorical artifact* output with justification, and *validation against human MOS* on every sequence. Wu et al. [24] is the closest antecedent (a systematic zero-shot IQA benchmark with human validation), but covers general distortions without full-reference codec comparison or artifact taxonomy. Trained VQA models miss the zero-shot/API and discrete-artifact axes; objective metrics miss interpretability; codec studies miss subjective validation. The combination is not merely additive: it is the artifact label validated against a controlled family comparison that reveals *blocking* as a one-sided, precision-1 signal for traditional codecs — a finding invisible to any single-axis prior work. This work occupies that intersection and reports its limits as clearly as its strengths.

3 METHOD

Our judge is a thin, training-free wrapper around a commercial VLM. The pipeline (Figure 1) has three stages.

(1) Aligned crop sampling. For each PVS we sample **2 frames** and extract **1 spatially aligned reference/reconstruction crop** (512 $\times$ 512) per frame, yielding **2 crop pairs per PVS** (1 pair in 12 of 216 PVS where only one frame was extractable at the target positions), at the display resolution (the reconstruction upsampled to the UHD source). Crop positions are deterministic, seeded by sequence name and frame index, so the choice of region is reproducible and not hand-picked. We send native-resolution crops rather than a downsampled full frame, because providers may downsample large images internally, which would risk erasing the fine texture blur we aim to detect; we treat this as a design choice rather than an ablated variable.

(2) Structured judging (anchored, two passes). Each judge receives the reference and reconstructed crop under a system prompt priming it as a perceptual assessor, and is queried in **two passes** returning JSON. The *score* pass asks for an integer 0–100 **anchored to the five-grade ITU-R BT.500 impairment scale** (imperceptible=100, perceptible-but-not-annoying=80, . . . , very-annoying=20). The *artifact* pass asks for the dominant label from the taxonomy (blocking, texture-blur, color-shift, ringing, none) with explicit per-label definitions, an instruction *not* to default to texture-blur, and a one-sentence justification. We adopted this design after a qualitative pilot on a subset of sequences: a single unanchored prompt collapsed nearly every label to *texture-blur* and visibly degraded rank agreement with MOS, whereas anchoring the score and decoupling the artifact pass recovered both correlation and label diversity. A controlled, held-out prompt comparison was not run; this design choice is therefore a researcher degree of freedom, as noted in Sec. 6. We enforce the schema via each provider’s structured-output mode; non-conforming responses are caught and excluded.

(3) Aggregation. A PVS yields multiple crops; we aggregate per-crop scores to a per-PVS value by averaging over its crops. The three models (**Claude (claude-opus-4-8)**, **GPT-5.5**, and **Gemini (gemini-3.5-flash)**) are evaluated independently and against each other, with returned model-version strings logged for reproducibility (commercial models drift).

4 EXPERIMENTAL SETUP

This section describes the corpus, the three judges and their baselines, and the statistical protocol applied throughout.

Dataset. AVT-VQDB-UHD-1-NVC [15, 16] (UHD-1, 6 source contents, neural DCVC-FM/RT + traditional AV1/VVC across quality levels, 216 PVS with per-PVS human MOS). We evaluate on *all* 216 PVS; **no new human study** is conducted. Larger compressed-VQA datasets exist (AIM-2024 [18], LEHA-CVQAD [6]) but lack the neural-vs-traditional MOS pairing central to our question. The following describes the three judges evaluated on this corpus and the baselines they are compared against.

Judges and input. Three frontier commercial models queried zero-shot: Claude (claude-opus-4-8), GPT-5.5, and Gemini (gemini-3.5-flash). Each receives the native-resolution 512 $\times$ 512 aligned crops described in Sec. 3, with the anchored two-pass prompt (ITU-R BT.500 score + artifact label). Exact model-version strings are logged; providers may still re-encode crops internally, which we treat as a limitation (Sec. 6). To isolate each predictor’s contribution, baselines are computed on the same input crops the judges receive.

Baselines. We recompute PSNR, SSIM, and LPIPS on the *same* 512 $\times$ 512 crops the judges see, so both operate on identical inputs. VMAF is kept at its native *full-video* level (it is a video metric; running it on isolated crops is not its intended operating mode) and taken directly from the dataset’s provided per-PVS scores (full-reference, as documented in [15]); we flag it as the one baseline with a whole-sequence advantage. We also report the dataset’s video-level PSNR/SSIM/LPIPS (Table 1 note). Comparing all predictors rigorously requires the following statistical protocol.

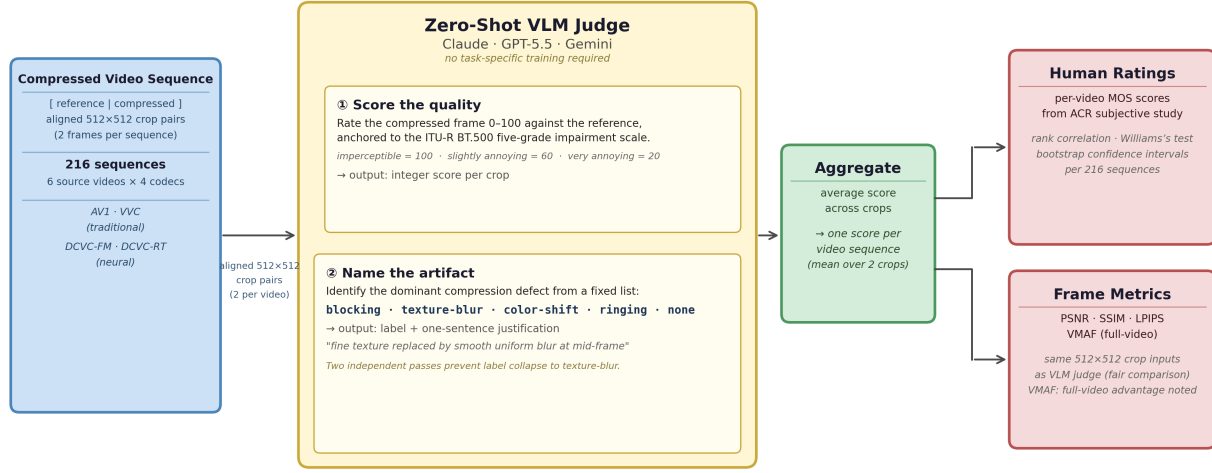


Figure 1: Zero-shot full-reference VLM judge pipeline. Aligned native-resolution crop pairs are scored by three commercial VLMs in two passes—ITU-R BT.500-anchored quality score and dominant-artifact label—and validated against human MOS and frame-level metrics.

Statistical analysis. Scores are aggregated to PVS level (mean over crops), matching the per-PVS MOS. Because the MOS distribution is non-normal (Shapiro–Wilk $p < 0.001$; Table 2), our **primary** criterion is the rank-based **SROCC** (Spearman Rank-Order Correlation Coefficient), with KROCC (Kendall Rank-Order Correlation Coefficient) reported alongside. PLCC and RMSE are also reported, but only after the ITU-T P.1401 five-parameter logistic mapping, with normality of residuals re-checked before any Pearson-based interpretation. Both pooled and **per-content** SROCC are reported to avoid the content confound. All correlations carry **95% confidence intervals** (CI): a content-level **bootstrap** ($B=2000$, resampling the 6 source contents with replacement — a non-parametric approach suited to the small content count) for SROCC, and a Fisher- z transform CI for PLCC.

To compare judges against VMAF, we use **Williams’s test** [9] for dependent correlations sharing MOS (Holm-corrected across four baselines), and a paired per-PVS absolute-error comparison (paired t -test and Wilcoxon signed-rank test, both reported because errors may not be normal). Significance: $\alpha=0.05$. Artifact labels carry no numeric ground truth and are analyzed distributionally, as described next.

Artifact analysis (RQ2). The dataset provides no artifact-type ground truth, so we analyze the *distribution* of each judge’s dominant-artifact label across the codec families and the agreement among the three judges, rather than accuracy against a human key. The discriminative question (whether artifact *type* separates traditional from neural codecs) is answered by the per-family label distribution (Fig. 4).

5 RESULTS

All numbers are over the full set of 216 PVS (6 contents $\times$ four codecs (AV1, VVC, DCVC-FM, DCVC-RT) $\times$ quality levels); judge scores are pooled over native-resolution crops.

5.1 RQ1: Correlation with MOS

Table 1 and Fig. 2 report agreement with human MOS. Under *matched conditions* (judges and objective metrics scored on identical 512x512 crops), the three VLM judges (SROCC 0.71–0.74) *outperform* every frame-level objective metric: PSNR (0.63), LPIPS (0.62), and SSIM (0.59). The gap is significant: under Williams’s test for dependent correlations, each judge beats each frame-level metric (e.g. Gemini vs. SSIM $t=+4.6$, $p < 10^{-5}$; vs. PSNR $t=+4.0$, $p < 10^{-4}$; vs. LPIPS $t=+13.9$, $p < 10^{-15}$; the weakest case, GPT-5.5 vs. PSNR, is still $p=5 \times 10^{-3}$). The judges trail only **VMAF** (0.91) — but VMAF is a *full-video*, trained metric that integrates the whole sequence, an advantage the single-crop judge does not have; that gap too is significant (Williams $t < -9$, $p < 0.001$; paired error test judge mean absolute error (MAE) 0.61–0.66 vs. VMAF 0.37, t -test and Wilcoxon $p < 0.001$). We report both tests because a Shapiro–Wilk test rejects normality of the MOS distribution ($p < 0.001$, Table 2). KROCC values confirm the same ordering (Gemini 0.553, Claude 0.540, GPT-5.5 0.523; VMAF 0.731; Table 1). The spread among the three judges (0.705–0.736) lies within the overlapping bootstrap confidence intervals (Table 1); no within-group ranking is statistically warranted. In summary, a zero-shot judge does *not* surpass full-video VMAF, but on *matched per-frame input* it extracts more quality-relevant signal than PSNR, SSIM, or LPIPS, with no training. For context, those metrics’ *video-level* scores are higher (PSNR 0.77, SSIM 0.85, LPIPS 0.72; Table 1 note): i.e., much of their apparent strength is whole-sequence integration, not per-frame perception. Two observations reinforce the result: **per-content** SROCC rises

Table 1: Agreement with human MOS over all 216 PVS (pooled). SROCC and KROCC are primary (rank-based; MOS is non-normal); 95% CI from content-level bootstrap on SROCC; PLCC after ITU-T P.1401 logistic mapping. VLM judges and frame-crop metrics use identical 512×512 inputs; VMAF uses full-video.

Predictor	SROCC	KROCC	95% CI	PLCC	SROCC _{cont}
<i>Full-video reference</i>					
VMAF	0.907	0.731	[0.88, 0.95]	0.911	0.940
<i>Frame-crop VLM judges (ours)</i>					
Gemini	0.736	0.553	[0.69, 0.83]	0.732	0.793
Claude	0.723	0.540	[0.64, 0.83]	0.719	0.811
GPT-5.5	0.705	0.523	[0.64, 0.80]	0.711	0.750
<i>Frame-crop objective metrics (same crops)</i>					
PSNR	0.627	0.444	[0.55, 0.74]	0.618	0.734
LPIPS*	0.623	0.447	[0.53, 0.77]	0.629	0.695
SSIM	0.585	0.415	[0.46, 0.72]	0.560	0.621

*LPIPS is a distance (negative SROCC); magnitude shown. For reference, at the dataset’s *video* level these metrics score higher (SROCC: PSNR 0.77, SSIM 0.85, LPIPS 0.72): much of their strength comes from integrating the whole sequence, which the single-crop judge does not see.

to 0.75–0.81 for the judges (above the frame metrics’ 0.62–0.73), so within a content they correctly rank quality levels; and the three judges show high agreement with one another (inter-model Spearman 0.88–0.95), so the verdicts reflect a systematic perceptual signal, not one vendor’s idiosyncrasy. The pattern holds on both codec families (Fig. 2); the scatter against MOS (Fig. 3) shows the judges’ spread relative to VMAF.

Stratified analysis by codec family. Table 3 repeats the correlation analysis of Table 1 separately for the traditional (AV1+VVC) and neural (DCVC-FM/RT) halves of the corpus ($n=108$ each), to establish whether the judges perform equally well in both domains. Three observations follow. **(i)** The overall ordering is preserved within each family: every judge outperforms every frame-level metric on traditional *and* on neural content, and trails only VMAF in both. **(ii)** All three judges are somewhat stronger on traditional codecs (SROCC 0.738–0.778) than on neural ones (0.688–0.706), but the content-level bootstrap CIs of the two subsets overlap for every judge, so the data do not support a significant within-judge family difference. **(iii)** The traditional-to-neural drop is modest for the judges (0.04–0.07 SROCC) and essentially absent for VMAF (0.912 vs. 0.899), whereas PSNR degrades markedly on neural content (0.682 to 0.558, a 0.12 drop) — consistent with the observation (Sec. 1) that signal-fidelity measures over-penalize neural texture re-synthesis that human viewers judge mildly. In summary, the zero-shot judges transfer across codec families with only a modest, statistically unconfirmed advantage on traditional content, and their margin over the frame-level metrics is, if anything, *wider* in the neural domain, where such an artifact-aware tool is most needed. Beyond correlation with MOS, we ask whether the judges’ artifact labels capture what a scalar score cannot.

Table 2: Shapiro–Wilk normality tests ($n=216$). MOS is non-normal, justifying the rank-based SROCC as the primary criterion; logistic-fit residuals are non-normal in four of the five cases, so PLCC-based claims are cross-checked with the non-parametric Wilcoxon test.

Variable	W	p	decision
MOS	0.934	< 0.001	non-normal
residual (VMAF)	0.973	< 0.001	non-normal
residual (SSIM)	0.944	< 0.001	non-normal
residual (Claude)	0.996	0.80	normal
residual (Gemini)	0.988	0.065	normal
residual (GPT-5.5)	0.974	< 0.001	non-normal

Table 3: Stratified agreement with human MOS by codec family (traditional AV1+VVC vs. neural DCVC-FM/RT; $n=108$ each). Protocol as in Table 1: 95% CI from content-level bootstrap on SROCC; PLCC after the ITU-T P.1401 logistic mapping refitted per subset; LPIPS magnitudes shown.

Predictor	SROCC	95% CI	KROCC	PLCC
<i>Traditional (AV1+VVC)</i>				
VMAF	0.912	[0.88, 0.96]	0.745	0.919
Gemini	0.778	[0.76, 0.84]	0.601	0.778
Claude	0.744	[0.66, 0.83]	0.561	0.746
GPT-5.5	0.738	[0.70, 0.82]	0.561	0.743
PSNR	0.682	[0.60, 0.78]	0.490	0.678
LPIPS	0.650	[0.57, 0.78]	0.471	0.641
SSIM	0.608	[0.47, 0.73]	0.437	0.601
<i>Neural (DCVC-FM/RT)</i>				
VMAF	0.899	[0.86, 0.95]	0.714	0.910
Gemini	0.706	[0.60, 0.85]	0.526	0.693
Claude	0.702	[0.62, 0.84]	0.527	0.695
GPT-5.5	0.688	[0.56, 0.81]	0.502	0.684
LPIPS	0.594	[0.46, 0.77]	0.424	0.597
PSNR	0.558	[0.47, 0.71]	0.394	0.562
SSIM	0.555	[0.35, 0.72]	0.393	0.588

5.2 RQ2: Artifact type and codec family

The dataset has no artifact-type ground truth, so we study the *distribution* of the judges’ dominant-artifact labels by codec family (Fig. 4). Two facts emerge. **(i) texture-blur dominates both families:** modern traditional codecs (AV1, VVC) deploy in-loop deblocking and, at the bitrates tested, degrade by smoothing fine texture much like the neural codecs do, so artifact *type* is *not* a broad family discriminator, contrary to the commonly reported lower-bitrate behavior of traditional codecs [19]. **(ii) When a structured artifact does appear, it is one-sided:** blocking is the majority label in 6 of the 108 traditional PVS (AV1+VVC; recall 5.6%) and in *zero* neural PVS (Fisher’s exact test, $p=0.014$), with the three judges unanimous on four of the six (two of three on the other two). Thus “blocking⇒traditional” is a high-precision, low-recall signal (0 false positives; 5.6% recall) that no scalar metric can provide, while blur is uninformative about the family. This nuance (that the discriminative power lies in the rare structured artifact, not in the modal label) is itself a finding about how modern codec families fail. We next examine what the

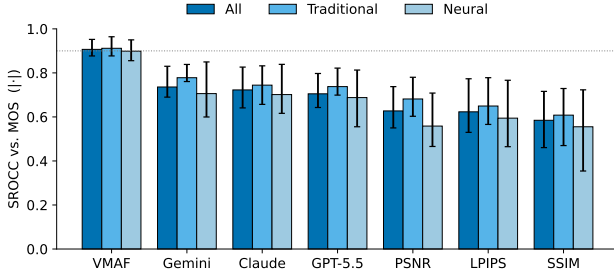


Figure 2: SROCC with human MOS per predictor (content-level bootstrap 95% CIs), overall and by codec family. VLM judges outperform frame-level metrics (PSNR, LPIPS, SSIM) and trail only full-video VMAF.

judges add beyond this artifact signal: interpretability, cross-vendor agreement, and the edges of their reliability.

5.3 RQ3: Interpretability beyond the score

Beyond out-correlating the frame-level metrics (RQ1), the judges’ distinctive value is an *auditable* verdict. For each PVS the judge returns a score *and* a one-sentence justification that localizes the degradation; VMAF returns only a scalar. Fig. 5 makes this concrete with two reconstructions at nearly identical perceptual quality (MOS 1.73 and 1.65) but different failure modes: the neural codec smooths fine bark and foliage texture (judges: *texture-blur*), while the traditional codec replaces the water surface with a blocky grid (judges: *blocking*). A scalar of equal magnitude conflates these two cases; the judge separates and explains them. The high inter-model agreement (Spearman 0.88–0.95) indicates these explanations are reproducible across independent vendors rather than vendor-specific artifacts. We therefore position the zero-shot VLM judge as a *complement* to trained metrics: when an explanation, an artifact label, or training-free operation on unseen codec types matters, it offers interpretability and zero-training generality while matching or exceeding classical frame-level metrics; full-video VMAF remains the choice when only the tightest scalar agreement with MOS is required. A full picture of the judges’ utility requires, however, an honest account of where they systematically fail.

5.4 Where the judge fails

Because all 216 PVS are scored, the failure modes can be characterized systematically rather than selectively. **(1) A monotonic resolution bias.** Splitting the signed error (judge minus MOS, in standard deviation (SD) units) by source resolution reveals that the judges *over-rate* low-resolution, upscaled clips and *under-rate* native 4K: +0.49 SD at $\leq 360p$ ($n=24$ PVS), +0.17 at 720p, −0.05 at 1080p, −0.22 at 2160p (correlation between resolution and signed error = −0.27, $p<0.001$). This suggests that VLMs are less sensitive than humans to the softness of upscaled low-resolution frames — the most systematic error source identified (the failure case in Fig. 5 is exactly this). **(2) Low blocking recall.** The artifact signal is high-precision but low-recall: blocking was never wrongly assigned to a neural codec (0/108 false positives), but it was detected in only

6 of 108 traditional PVS (recall 5.6%) — most degraded traditional clips are still labeled *texture-blur*, so the judges *miss* blocking far more often than they *confuse* it. **(3) Residual label disagreement.** On 3 of 216 PVS the three judges produced three different dominant labels (typically blur vs. color-shift on near-threshold clips); the remaining 213 (98.6%) reach majority agreement (≥ 2 of 3 on the same label), consistent with the high inter-judge agreement already reported in Sec. 5 (0.88–0.95). We report these explicitly as limitations of the method.

6 DISCUSSION AND LIMITATIONS

Frame-level judge vs. a video-level reference. To ensure the judge-vs-metric comparison uses identical inputs, PSNR/SSIM/LPIPS are recomputed on the *same* crops as the judges; only VMAF remains *video-level* (it is a video metric). Two asymmetries remain. First, MOS is collected on full videos while the judges (and the frame-crop metrics) see sampled frames, so *temporal* artifacts (flicker, inter-frame inconsistency, judder) are not directly observed: results are frame-level FR quality pooled to the sequence, not full temporal VQA. Second, VMAF’s whole-sequence view is a genuine advantage we do not control for; a fair judge-vs-VMAF contrast would require a temporal judge over frame sequences, which we leave as future work.

Provider-side image handling. Although we send native-resolution crops to avoid whole-frame downscaling, each provider may still re-encode or resize internally; we cannot observe the pixels actually ingested. The cross-provider comparison is therefore between three black-box pipelines, not three identical visual front-ends.

Absolute scoring. Consistent with prior reports [26] that VLMs rank better than they assign calibrated absolute scores, our primary evidence is rank correlation (SROCC/KROCC); PLCC is reported only after the standard logistic mapping. A pairwise (Bradley–Terry) judge is a natural extension if absolute calibration is needed.

Possible data exposure. The dataset and its MOS are publicly available and could appear in a model’s training data, which would inflate “zero-shot” correlation. We cannot fully exclude this; we note the dataset is recent relative to the models and that memorizing per-PVS numeric MOS is unlikely, but treat the absolute numbers with this caveat.

Single dataset and codec set. All results come from one dataset; strong claims are scoped to its UHD content and the two DCVC variants. We did not run a second dataset, so generalization to other content, resolutions, and neural codecs remains open — a deliberate scope choice given the lack of comparable neural-vs-traditional MOS elsewhere.

Model drift and determinism. API models are versioned and not guaranteed deterministic; we query at temperature 0 where supported (OpenAI and Gemini; the Claude API does not expose this parameter and returns an error if it is sent) and log the exact model-version string each API returns. To verify that single-sample scoring ($R=1$) is adequate, we re-queried the least costly judge (Gemini) *three times* over all PVS: per-crop scores are stable, with a mean standard deviation 0.13 on the 0–100 scale (median 0, 90th

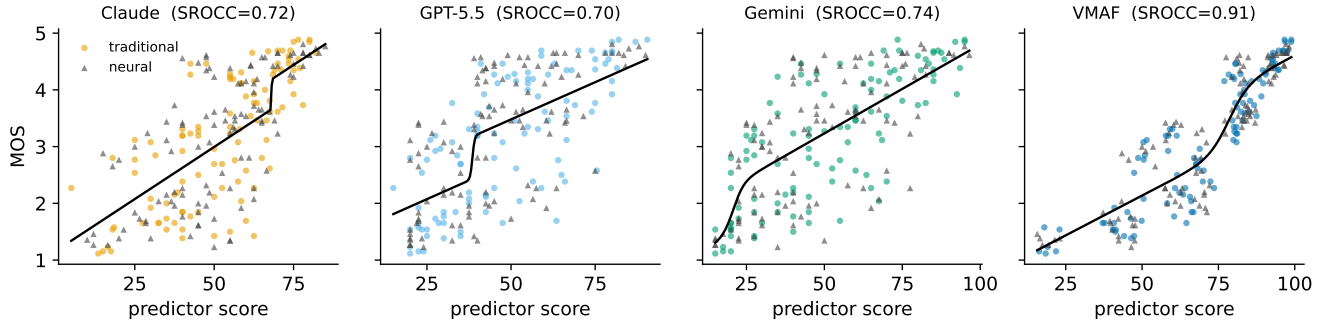


Figure 3: Per-PVS predictor score vs. human MOS with the ITU-T P.1401 logistic fit; markers distinguish traditional (circles) and neural (triangles) codecs. The judges track MOS but with a wider spread than VMAF.

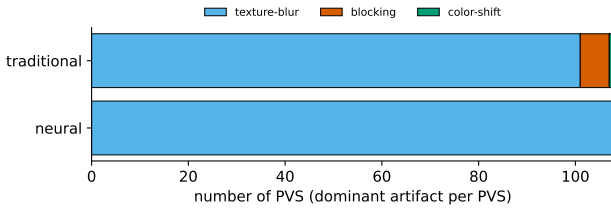


Figure 4: Dominant artifact per PVS (majority over judges) by codec family. Texture-blur dominates both; blocking occurs only for traditional codecs.

pct. 0), and a test–retest SROCC of 0.998 between repetitions. Intra-model variance is therefore negligible at temperature 0; the residual reproducibility risk is provider-side model drift, not sampling noise.

Prompt selection. Zero-shot results depend on the prompt. We did *not* hold out a separate prompt-selection set: our anchored two-pass design was chosen on pilot data because an un-anchored prompt under-performed (Sec. 3). This is a form of researcher degrees of freedom; a differently engineered prompt could shift the absolute numbers, and the prompt should be regarded as part of the method, not a neutral given. A systematic prompt-sensitivity study — several semantically equivalent prompt variants evaluated on a held-out validation split — is a necessary step before any production use of the approach.

Cost, latency, and accuracy scope. API inference is not cheaper than VMAF, and the judges do not surpass it on full-video correlation; we quantify the gap in order-of-magnitude terms. Each judge issues two structured calls (score pass and artifact pass) per crop pair, i.e., 840 calls per judge over the 420 crop pairs and roughly 2,500 calls for the full three-judge benchmark; every call uploads two 512×512 crops and returns a short JSON object. Per-call latency is on the order of seconds, so a full 216-PVS audit completes in tens of minutes to a few hours depending on request parallelism — against milliseconds per frame for locally computed PSNR/SSIM and faster-than-real-time VMAF. At the commercial list prices in effect during our runs, a single call costs a fraction of a US cent on the flash-priced tier (Gemini 3.5 Flash) and single-digit US cents

on the frontier tiers, placing the complete three-judge evaluation on the order of tens of US dollars — negligible at research scale, but two to three orders of magnitude more expensive per frame than local objective metrics, and subject to provider price changes. The judge is therefore not positioned as a bulk quality-monitoring replacement: the value it adds is *interpretability* — an artifact label and a justification sentence for targeted triage — not marginal accuracy gains over an already strong metric.

Why do general-purpose VLMs capture perceptual quality? We offer a mechanistic reading, explicitly as intuition rather than as an established account. Web-scale image–text pretraining exposes a VLM to a vast number of naturally captioned degraded images — photographs described as “blurry” or “blocky”, compression artifacts discussed by name — so the visual encoder plausibly learns representations in which common degradation patterns are both discriminable and *linked to their names*; Q-Bench [22] documents precisely this latent low-level sensitivity in general-purpose LMMs. Instruction tuning then allows a prompt to bind that latent sensitivity to a rubric: the ITU-anchored scale and the fixed artifact taxonomy act as a lightweight, inference-time calibration layer, consistent with the finding of Q-Align [23] that discrete text-defined levels outperform raw numeric regression. Specialized VQA models (e.g., CAMP-VQA [20], LMM-VQA [5]) instead optimize a regression objective against MOS on curated quality datasets, which yields tighter scalar calibration on in-distribution content but ties the model to the distortions and label space seen during training. The zero-shot judge inverts this trade-off: weaker absolute calibration (the resolution bias of Sec. 5) in exchange for transferable perceptual concepts and the ability to name what it sees.

Interpretability validated qualitatively. The usefulness of the justifications is demonstrated by example (Fig. 5) and by cross-vendor consistency, not by a user study: we did not measure whether the explanations improve or accelerate the decisions of practicing codec engineers. A controlled human evaluation of explanation usefulness (e.g., engineers triaging quality regressions with and without the artifact labels) is explicit future work.

Commercial-only models. All three judges are closed, commercial APIs (Claude, GPT-5.5, Gemini). Whether open-weight models (e.g., LLaVA [14], InternVL [3], Qwen-VL [1]) would match this performance without the same infrastructure costs is unknown;

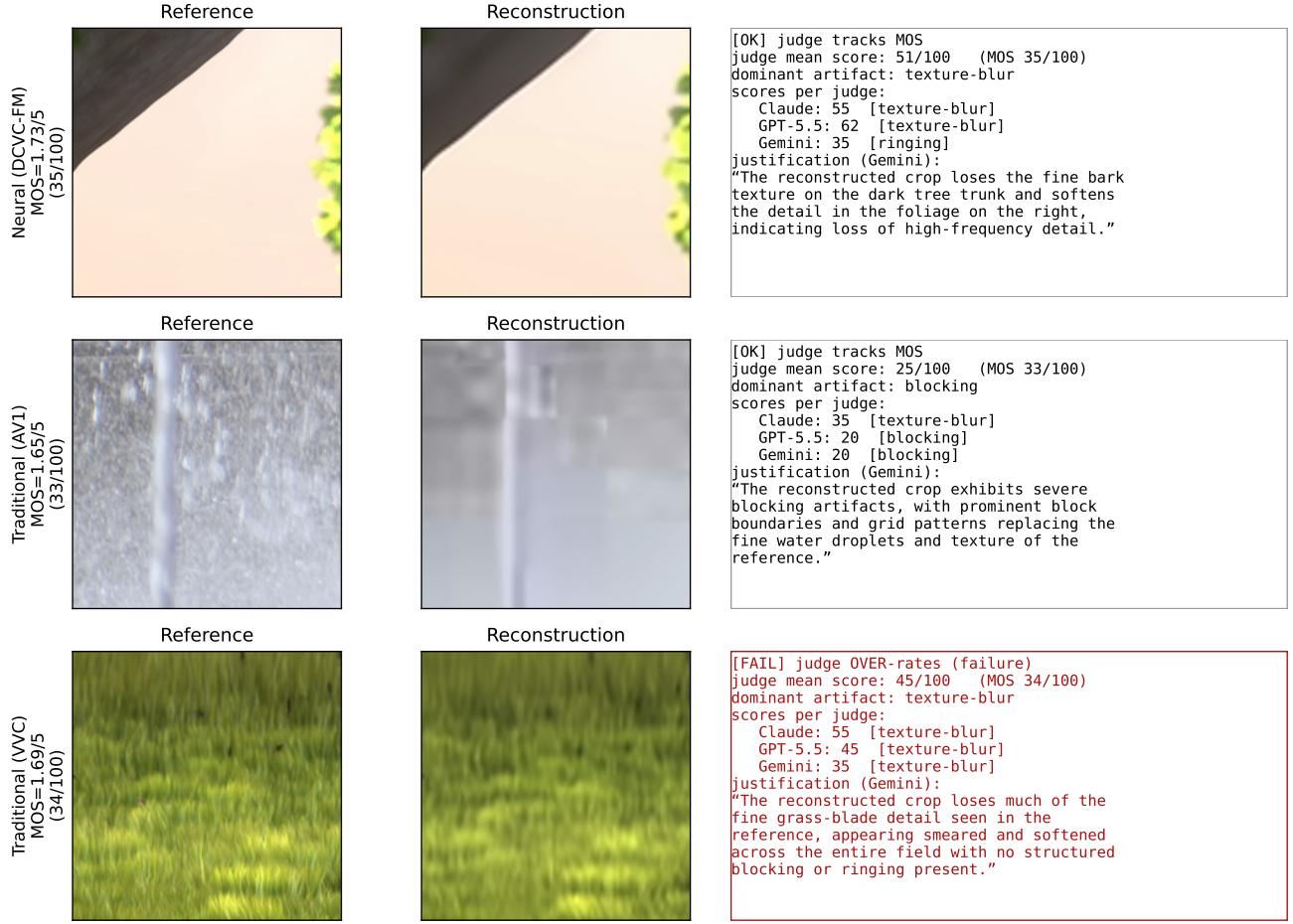


Figure 5: Illustrative cases (not a substitute for the full-216-PVS results above). Top/middle: judge tracks MOS—*texture-blur* on a neural codec and *blocking* on a traditional one, a distinction scalars cannot express. Bottom (failure): a low-resolution VVC clip the judges over-rate (~60 vs. MOS 34).

our findings may not generalize to models outside the frontier commercial tier, and open-source replication of the full pipeline is left as future work. The released evaluation suite (prompts, schema, scoring scripts) makes an open-weight replication under the identical protocol straightforward.

7 CONCLUSION AND FUTURE WORK

We studied whether off-the-shelf vision-language models can act as zero-shot, full-reference quality judges for compressed video that not only outperform classical frame-level metrics but are also *interpretable* — naming the dominant artifact and justifying the verdict, a capability absent from PSNR/SSIM/VMAF/LPIPS and from trained neural VQA. Using three commercial VLMs and validating against human MOS on a dataset spanning traditional and neural codecs, with native-resolution crops and a rigorous statistical protocol, we scoped the contribution: on matched per-frame input the judges outperform PSNR/SSIM/LPIPS and trail only full-video VMAF; their

value is interpretability and zero-training generality, not accuracy or cost. We also reported their systematic failures (a resolution bias; low blocking recall) rather than only their successes.

Future work includes a *temporal* judge (frame sequences rather than still crops), a pairwise (Bradley-Terry) formulation for calibrated absolute scores, an open-weight replication of the pipeline (LLaVA, InternVL, Qwen-VL), a human study of whether the artifact labels and justifications measurably assist codec engineers, and using the artifact label to *act* — steering post-processing quality-enhancement filters toward the specific degradation present, or driving artifact-aware adaptive codec selection.

ACKNOWLEDGEMENTS

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior — Brasil (CAPES) — Finance Code 001.

ETHICAL CONSIDERATIONS

This work did not require review by a Research Ethics Committee under Resolution 510/2016 of the Brazilian National Health Council (CEP/CONEP system), as it involves no human participants and no personal data. All experiments query commercial AI services with video frames from a public corpus. The subjective quality scores (MOS) used as ground truth come exclusively from the public dataset AVT-VQDB-UHD-1-NVC [15], whose subjective evaluations were collected by the original dataset authors under their own institutional protocols; no new human study was conducted by the authors of this paper. No personally identifiable information was collected, processed, or stored, and no privacy-regulation concerns apply. The authors declare no conflicts of interest.

GENERATIVE AI DISCLOSURE

A generative AI assistant (large language model) was used to support *writing and revision* of this manuscript. It was *not* used to generate research ideas, data, experimental results, or analyses: the study design, experiments, all reported numbers, statistical procedures, and conclusions are the sole responsibility of the authors, verified and fully owned by them. (This is distinct from the vision–language models studied as the *method*, described in Sec. 3.)

REFERENCES

- [1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. arXiv preprint arXiv:2308.12966.
- [2] Benjamin Bross, Ye-Kui Wang, Yan Ye, Shan Liu, Jianle Chen, Gary J. Sullivan, and Jens-Rainer Ohm. 2021. Overview of the Versatile Video Coding (VVC) Standard and its Applications. *IEEE Transactions on Circuits and Systems for Video Technology* 31, 10 (2021), 3736–3764. doi:10.1109/TCSVT.2021.3101953
- [3] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024. InternVL: Scaling up Vision Foundation Models and Aligning for Generic Visual-Linguistic Tasks. In *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*. 24185–24198. doi:10.1109/CVPR52733.2024.02283
- [4] Keyan Ding, Kede Ma, Shiqi Wang, and Eero P. Simoncelli. 2022. Image Quality Assessment: Unifying Structure and Texture Similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)* 44, 5 (2022), 2567–2581. doi:10.1109/TPAMI.2020.3045810
- [5] Qihang Ge, Wei Sun, Yu Zhang, Yunhao Li, Zhongpeng Ji, Fengyu Sun, Shangling Jui, Xiongkuo Min, and Guangtao Zhai. 2025. LMM-VQA: Advancing Video Quality Assessment with Large Multimodal Models. *IEEE Transactions on Circuits and Systems for Video Technology* 35, 11 (2025), 11083–11096. doi:10.1109/tcsvt.2025.3571788
- [6] Aleksandr Gushchin et al. 2025. LEHA-CVQAD: Dataset to Enable Generalized Video Quality Assessment of Compression Artifacts. In *Proc. 33rd ACM Int. Conf. on Multimedia (ACM MM)*. 13405–13412. doi:10.1145/3746027.3758303
- [7] Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haoan Chen, Abhranil Chandra, Ziyang Jiang, Aaran Arulraj, Kai Wang, Quy Duc Do, Yuansheng Ni, Bohan Lyu, Yaswanth Narasupalli, Rongqi Fan, Zhiheng Lyu, Yuchen Lin, and Wenhui Chen. 2024. VideoScore: Building Automatic Metrics to Simulate Fine-grained Human Feedback for Video Generation. In *Proc. Conf. on Empirical Methods in Natural Language Processing (EMNLP)*. 2105–2123. doi:10.18653/v1/2024.emnlp-main.127
- [8] ITU-R. 2023. *Recommendation BT.500-15: Methodologies for the Subjective Assessment of the Quality of Television Images*. Technical Report. International Telecommunication Union. <https://www.itu.int/rec/R-REC-BT.500>
- [9] ITU-T. 2020. *Recommendation P.1401: Methods, Metrics and Procedures for Statistical Evaluation, Qualification and Comparison of Objective Quality Prediction Models*. Technical Report. International Telecommunication Union. <https://www.itu.int/rec/T-REC-P.1401>
- [10] ITU-T. 2023. *Recommendation P.910: Subjective Video Quality Assessment Methods for Multimedia Applications*. Technical Report. International Telecommunication Union. <https://www.itu.int/rec/T-REC-P.910>
- [11] Zhaoyang Jia, Bin Qi, Jiahao Li, Bin Li, Ruiqi Xie, Houqiang Li, and Yan Lu. 2025. Towards Practical Real-Time Neural Video Compression (DCVC-RT). In *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*. 12543–12552. doi:10.1109/cvpr52734.2025.01170
- [12] Jiahao Li, Bin Li, and Yan Lu. 2024. Neural Video Compression with Feature Modulation (DCVC-FM). In *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*. 26099–26108. doi:10.1109/cvpr52733.2024.02466
- [13] Zhi Li, Anne Aaron, Ioannis Katsavounidis, Anush Moorthy, and Megha Manohara. 2016. Toward a Practical Perceptual Video Quality Metric (VMAF). Netflix Technology Blog. <https://netflixtechblog.com/toward-a-practical-perceptual-video-quality-metric-653f208b9652>
- [14] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, Vol. 36. 34892–34916. <https://neurips.cc/virtual/2023/poster/70080>
- [15] Rakesh Rao Ramachandra Rao et al. 2025. Evaluating Video Quality Metrics for Neural and Traditional Codecs using 4K/UHD-1 Videos (AVT-VQDB-UHD-1-NVC). arXiv preprint arXiv:2511.00969. Dataset used in this work; preprint from TU Ilmenau.
- [16] Rakesh Rao Ramachandra Rao, Steve Göring, Patrick Vogel, Nicolas Pachatz, Juan Villamar Villarreal, Werner Robitzka, Peter List, Bernhard Feiten, and Alexander Raake Müller. 2019. Adaptive Video Streaming with Current Codecs and Formats: Extensions to Parametric Video Quality Models (AVT-VQDB-UHD-1). In *Proc. IS&T Electronic Imaging: Image Quality and System Performance XVI*. doi:10.2352/issn.2470-1173.2019.10.iqsp-314
- [17] Reza Rassool. 2017. VMAF Reproducibility: Validating a Perceptual Practical Video Quality Metric. In *Proc. IEEE Int. Symp. on Broadband Multimedia Systems and Broadcasting (BMSB)*. doi:10.1109/BMSB.2017.7986143
- [18] Maksim Smirnov et al. 2024. AIM 2024 Challenge on Compressed Video Quality Assessment: Methods and Results. In *Computer Vision – ECCV 2024 Workshops (Lecture Notes in Computer Science, Vol. 15632)*. 212–228. doi:10.1007/978-3-031-91856-8_13
- [19] Leandro Tavares, Ruhan Conceição, Victor Costa, Luciano Agostini, Marcelo Porto, and Guilherme Corrêa. 2025. Traditional vs. Neural Video Codecs: Compression Efficiency, Visual Artifacts, and Quality Analysis Beyond PSNR. In *Proc. Brazilian Symp. on Multimedia and the Web (WebMedia)*. SBC, 446–454. doi:10.5753/webmedia.2025.16011
- [20] Xinyi Wang, Angeliki Katsenou, Junxiao Shen, and David Bull. 2026. CAMP-VQA: Caption-Embedded Multimodal Perception for No-Reference Quality Assessment of Compressed Video. In *Proc. IEEE/CVF Winter Conf. on Applications of Computer Vision (WACV)*. 2042–2051. <https://ieeexplore.ieee.org/document/11492275>
- [21] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. 2004. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612. doi:10.1109/TIP.2003.819861
- [22] Haoning Wu, Zicheng Zhang, Erli Zhang, Chaofeng Chen, Liang Liao, Annan Wang, Chunyi Li, Wenxiu Sun, Qiong Yan, Guangtao Zhai, and Weisi Lin. 2024. Q-Bench: A Benchmark for General-Purpose Foundation Models on Low-Level Vision. In *Proc. Int. Conf. on Learning Representations (ICLR), Spotlight*. <https://openreview.net/forum?id=0V5TVt9bk0>
- [23] Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Wei Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. 2024. Q-Align: Teaching LLMs for Visual Scoring via Discrete Text-Defined Levels. In *Proc. Int. Conf. on Machine Learning (ICML), PMLR vol. 235*. <https://proceedings.mlr.press/v235/wu24ah.html>
- [24] Tianhe Wu, Kede Ma, Jie Liang, Yujiu Yang, and Lei Zhang. 2024. A Comprehensive Study of Multimodal Large Language Models for Image Quality Assessment. In *Computer Vision – ECCV 2024 (Lecture Notes in Computer Science, Vol. 15132)*. Springer Nature Switzerland, 143–160. doi:10.1007/978-3-031-72904-1_9
- [25] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*. 586–595. doi:10.1109/CVPR.2018.00068
- [26] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*. <https://openreview.net/forum?id=ucCHPGdlao>
- [27] Hanwei Zhu, Xiangjie Sui, Baoliang Chen, Xuelin Liu, Peilin Chen, Yuming Fang, and Shiqi Wang. 2024. 2AFC Prompting of Large Multimodal Models for Image Quality Assessment. *IEEE Transactions on Circuits and Systems for Video Technology* 34, 12 (2024), 12873–12878. doi:10.1109/tcsvt.2024.3434999