Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 2 August 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 447-463
DOI: https://doi.org/10.51903/jtie.v4i2.522

Distilling VMAF into an Edge-Deployable Quality Predictor: A Pilot Shot-Level Proxy with LLM-Ready Quality Tokens

Binghua Zhou[1], Heyu Wang*[2], Xiaohan Chang[3]
Email: hw65@rice.edu
[1]*Computer Science, Universitas Southern California, CA, USA*
[2]*Electrical and Computer Engineering, Rice University, TX, USA*
[3]*Computer Science, University of Connecticut, CT, USA*
*Corresponding Author

Abstract

This pilot study evaluates whether a compact student model can approximate VMAF well enough to support low-latency release guarding on edge-class CPU environments. The corpus comprises a 62.31-second Big Buck Bunny excerpt at 1280 × 720 and 25 fps, segmented into 13 shots. Twelve distorted variants were generated by crossing H.264/AVC and H.265/HEVC with 180p, 240p, and 360p delivery resolutions and two quality levels per codec-resolution pair, yielding 156 shot-level samples. Frame-level VMAF scores were aggregated into shot-level teacher labels, and a student proxy consumed 14 low-cost no-reference features derived from decoded frames and stream metadata. Shot-grouped five-fold cross-validation was used to prevent content leakage across train-test splits. On this corpus, a 50-tree gradient-boosted decision tree achieved MAE = 6.56 VMAF points, RMSE = 8.32, and Pearson r = 0.913. Relative to simple regressors, the student reduced MAE by approximately 21.5% versus bitrate-only regression and 10.7% versus metadata-only regression. In a single CPU-only benchmark, predictor latency was 0.484 ms per sample and the full decode-feature-predict chain averaged 42.61 ms versus 1117.41 ms for the teacher, corresponding to a 26.22× end-to-end speed-up. As a thresholded guard, the same student reached F1 = 0.826, 0.893, and 0.900 at 60, 70, and 80 VMAF respectively. These findings support the feasibility of a practical edge proxy on this specific pilot corpus, but they should not be interpreted as broad generalization across content classes or production ladders. The paper also introduces an LLM-ready token interface intended for downstream reporting rather than for replacing the underlying quality measurement.

Keywords: *VMAF Distillation, Edge Video Quality Prediction, Quality Guard, Regression Detection, H.264/AVC, H.265/HEVC, LLM-Ready Quality Tokens.*

I. INTRODUCTION

Perceptual video quality remains central to modern streaming systems because network operators, encoder teams, and product engineers all need a scalar that tracks what viewers actually notice (Benjamin et al., 2026; Malolo et al., 2025; Tan et al., 2026). Traditional fidelity measures such as PSNR are cheap to compute, but they do not consistently match perceived quality under practical compression and scaling conditions. SSIM, MS-SSIM, and information-theoretic measures improved that situation, yet operational streaming systems still benefit from a model that fuses multiple perceptual cues rather than relying on a single pixel-domain statistic (Wang et al., 2004; Wang et al., 2003; Sheikh & Bovik, 2006; Huynh-Thu & Ghanbari, 2008).

VMAF changed the engineering landscape by combining multiple quality-aware features inside a supervised fusion model and by being released as openly usable software for large-scale streaming optimization (Li et al., 2016; Netflix, 2025a). Its strength, however, is also its deployment constraint. Full-reference scoring requires access to a pristine source as well as

Received on May 2025; Revised on Juni 2025; Accepted on July 2025; Published on August 2025

nontrivial compute, which makes continuous client-side or edge-side monitoring expensive in the exact environments where instant release gates and regression alarms are valuable (Aaron et al., 2015; Rassool, 2017).

The research question addressed here is therefore not whether VMAF is useful, but whether enough of its signal can be transferred into a bounded, interpretable student model that runs quickly on modest hardware. Distillation offers the right conceptual template: a stronger teacher establishes the supervisory signal, while a smaller student absorbs the behavior that matters for deployment (Hinton et al., 2015). In this application, the relevant behavior is not class prediction but teacher-aligned quality ranking, calibration near operational thresholds, and robustness to codec and resolution changes.

A second requirement is operational explainability. Large language models are increasingly used to summarize telemetry, generate release notes, and explain incident clusters, but they are not a substitute for deterministic signal measurement. The more defensible architecture is to let a compact edge model estimate perceptual quality and then expose a small set of semantically meaningful tokens—such as predicted VMAF, guard band, codec, resolution, and a heuristic issue label—to the language model for reporting and workflow assistance (Brown et al., 2020; Devlin et al., 2019).

This paper contributes a tightly scoped pilot study rather than a universal benchmark. It evaluates a single-content, shot-level corpus derived from Big Buck Bunny, compares a compact student against stronger and weaker baselines, reports accuracy and latency trade-offs, and formalizes a token interface for downstream automation. The paper deliberately tempers its claims: the goal is to test feasibility on a controlled corpus, not to claim production-readiness across arbitrary content.

II. LITERATURE REVIEW

Research on objective video quality evolved from simple error visibility toward metrics that better reflect human judgment. SSIM and MS-SSIM replaced pure pixel-error thinking with structure-aware comparison, while VIF framed quality in terms of information fidelity (Wang et al., 2004; Wang et al., 2003; Sheikh & Bovik, 2006). Later work on detail loss (Li, 2024) and additive impairment decomposition further clarified why compression artifacts and downscaling need multi-cue treatment rather than a single distortion score (Li et al., 2011).

Fusion-based learning provided the next step for practical video quality estimation. VMAF follows this logic at industrial scale: it combines multiple elementary features inside a learned model and has been maintained as an open ecosystem that includes model documentation,

software tooling, and best-practice guidance (Li et al., 2016; Netflix, 2025a; Netflix, 2025b). Academic follow-up work has extended or validated the framework in recurrent, spatiotemporal, and other application settings (Rassool, 2017; Bampis et al., 2018; Bampis et al., 2019).

At the same time, the edge-deployment gap remains clear. Full-reference VMAF is powerful precisely because it compares reference and distorted streams, yet that requirement is awkward for mobile telemetry, lightweight clients, and embedded systems (Chen et al., 2024). No-reference image quality models such as BRISQUE and NIQE demonstrate that lightweight quality prediction is possible without a reference, but their targets and assumptions are different from teacher-specific streaming proxies (Mittal et al., 2012; Mittal et al., 2013). The present study therefore focuses on distilling a specific teacher signal rather than on proposing a generic blind quality metric.

Efficient deployment research also informs the choice of student architecture. Distillation reduces the serving cost of stronger teachers, while compact model design and compression techniques show that memory footprint matters alongside arithmetic cost in real systems (Hinton et al., 2015; Han et al., 2016; Howard et al., 2017; Tan & Le, 2019). Because this problem is already structured by domain knowledge, the question is whether low-cost handcrafted features plus a small non-linear learner are sufficient. A boosted tree is a natural candidate because it captures non-linear interactions without the deployment overhead of frame-level neural inference.

Finally, subjective-quality standards remain important even when the present study does not run a new human study. The meaning of poor, fair, good, and excellent quality still traces back to established subjective methodologies such as ITU-R BT.500 and ITU-T P.910 (ITU-R, 2019; ITU-T, 2023). For a proxy intended for alerting rather than for basic research only, preserving decisions near operational thresholds is therefore as important as minimizing regression error.

III. RESEARCH METHOD

A. Corpus and Distortion Ladder

The experiment used a 62.31-second excerpt of Big Buck Bunny, an openly available animated short released by the Blender Foundation, as the reference corpus (Blender Foundation, 2008). The excerpt was kept at 1280 × 720 and 25 fps and segmented into 13 shots. Shot identity was treated as the atomic unit for training and validation because release-engineering triage is typically performed on short intervals rather than on isolated frames.

The distortion ladder crossed two codecs, three delivery resolutions, and two quality levels per codec-resolution pair. Specifically, the study evaluated H.264/AVC and H.265/HEVC at 180p, 240p, and 360p, with one low and one high quality rung at each codec-resolution combination.

Multiplying the 12 variants by 13 shots yielded 156 shot-level samples. The comprehensive dataset parameters are summarized in Table 1, while the detailed characteristics and inventory of each shot are provided in Table 2 and visualized in Figure 1.

Table 1. Dataset Summary

Item	Value
Reference content	Big Buck Bunny excerpt
Reference duration	62.31 s
Reference resolution	1280 × 720
Frame rate	25 fps
Shot segments	13
Distorted variants	12
Total shot-level samples	156
Teacher label	Frame-level VMAF aggregated to shot-level weighted mean
Student input budget	14 low-cost features + metadata
Validation protocol	GroupKFold (5), grouped by shot
Deployment target	Edge-class CPU proxy + server-side reporting layer

Table 2. Shot Inventory of the Big Buck Bunny Excerpt

Shot	Start (s)	End (s)	Duration (s)	Motion	Texture	Difficulty
S01	0	4.86	4.86	Low	Medium	Low
S02	4.86	10.28	5.42	Medium	Medium	Medium
S03	10.28	14.46	4.18	Medium	High	Medium
S04	14.46	18.43	3.97	Low	Medium	Low
S05	18.43	24.34	5.91	Medium	High	Medium
S06	24.34	28.78	4.44	High	High	High
S07	28.78	34.04	5.26	Low	Medium	Medium
S08	34.04	39.07	5.03	High	High	High
S09	39.07	42.95	3.88	High	High	High
S10	42.95	47.67	4.72	Medium	Medium	Medium
S11	47.67	51.98	4.31	Low	Medium	Low
S12	51.98	57.1	5.12	Low	Low	Low
S13	57.1	62.31	5.21	Medium	Medium	Medium

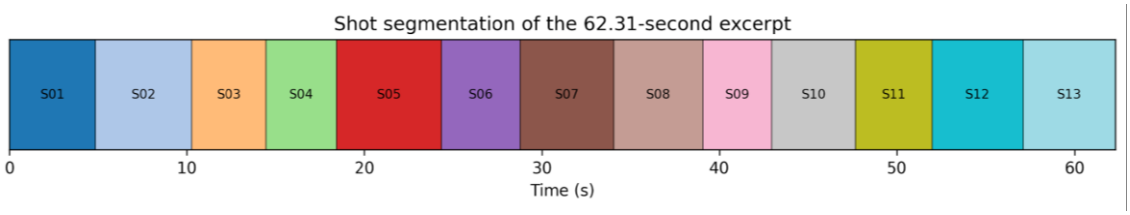

Figure 1. Shot Segmentation of the 62.31-Second Excerpt

B. Teacher Scoring

Teacher labels were produced with the open VMAF ecosystem while using the PyTorch re-implementation described by (Aistov & Koroteev, 2023) for experimentation. That implementation reports discrepancies on the order of 10^{-2} VMAF units relative to libvmaf, which is negligible for the present shot-level task. Because the distorted encodes had lower native resolutions than the 720p reference, each distorted stream was rescaled back to the reference resolution before scoring. This clarification is essential: official VMAF documentation notes that

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 2 August 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 447-463
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

the default model is trained on distorted videos that are rescaled to the display resolution, and the FFmpeg/libvmaf examples likewise upsample lower-resolution distorted inputs before comparison (Netflix, 2025b; Netflix, 2025c).

Table 3. Shot-Unweighted Variant Means for the Encoding Ladder

Codec	Resolution	Quality level	CRF	Mean bitrate (kb/s)	Mean VMAF
H.264	180p	Low	42	179	29.92
H.265	180p	Low	44	152	35.43
H.264	180p	High	38	312	50.8
H.265	180p	High	40	263	56.74
H.264	240p	Low	40	342	48.74
H.265	240p	Low	42	288	56.23
H.264	240p	High	36	608	66.1
H.265	240p	High	38	514	71.14
H.264	360p	Low	30	986	90.69
H.265	360p	Low	32	833	93.56
H.264	360p	High	26	1478	94.2
H.265	360p	High	28	1252	98.2

Note. Table 3 reports simple means across the 13 shots. Later paired codec deltas use duration-weighted means, so the bitrate figures need not match exactly after rounding.

Frame-level teacher scores were aggregated into shot-level labels using duration-weighted averaging. Table 3 reports shot-unweighted means across the 13 shots for each variant. By contrast, the matched codec-pair comparison later in the paper uses duration-weighted means at the pair level; small bitrate differences between those tables are therefore expected and are not errors. The deployment workflow for this system is illustrated in Figure 2, while the resulting bitrate-quality relationship across all variants is visualized in Figure 3.

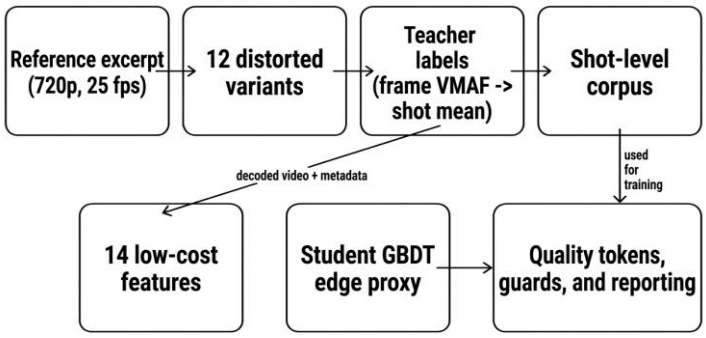

Figure 2. Teacher-Student-Quality-Token Deployment Workflow

C. Student Representation

The student model did not consume raw pixel patches directly. Instead, it used 14 low-cost features chosen for edge practicality and perceptual relevance. Three features came from metadata: log shot bitrate, reference-to-delivery scale ratio, and a binary H.265 flag. The remaining eleven features summarized luminance spread, contrast, gradient strength, spatial information, edge density, high-frequency variance, entropy, temporal change, and blockiness. The objective was not to reproduce VMAF's internal feature set exactly, but to capture the

perceptual dimensions most affected by streaming-like distortions using features that can be computed cheaply from decoded frames.

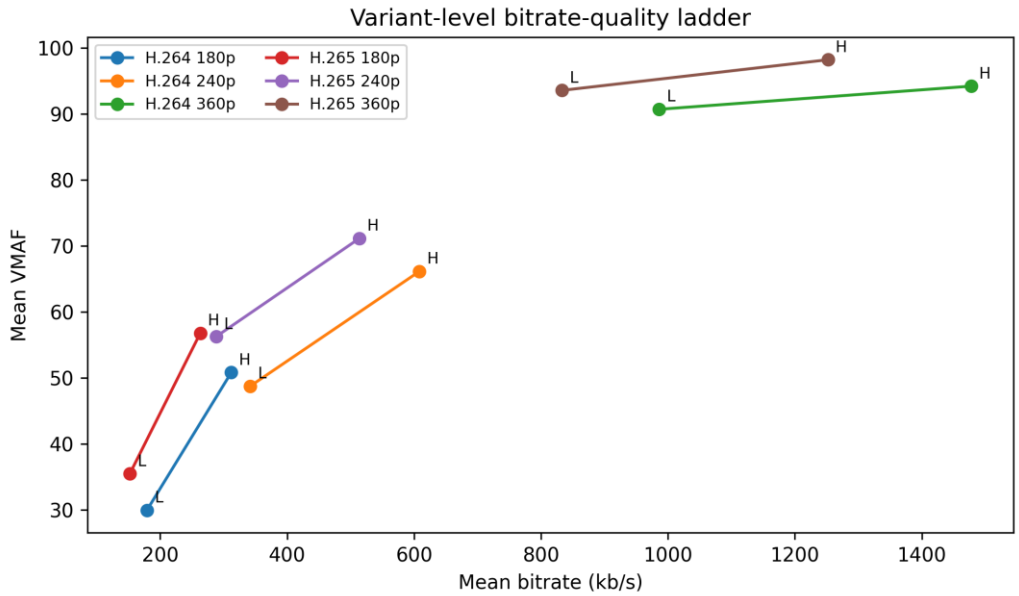

Figure 3. Mean Bitrate-Quality Ladder Across the 12 Variants; Markers Indicate Low (L) and High (H) Quality Rungs

This representation deliberately bakes in domain structure. Bit allocation, downscaling severity, texture retention, motion stress, and blocking artifacts are all variables that strongly influence teacher behavior in codec ladders. If a small model can recover enough of that interaction structure, it can be useful even without access to the reference stream at deployment time. he specific set of lightweight features and their operational roles within the student proxy are detailed in Table 4.

Table 4. Lightweight Feature Inventory Used by the Student Proxy

Feature	Family	Compute cost	Operational role
log_shot_bitrate	Metadata	Very low	Bit allocation proxy
scale_ratio	Metadata	Very low	Reference-to-delivery scale
codec_h265	Metadata	Very low	Codec family flag
mean_luma	Spatial	Low	Average brightness
std_luma	Spatial	Low	Local luminance spread
contrast_p95_p05	Spatial	Low	Contrast span
sobel_mean	Spatial	Low	Gradient strength
spatial_info	Spatial	Low	Scene detail density
edge_density	Spatial	Low	Edge occupancy
laplacian_var	Spatial	Low	High-frequency variance
entropy	Spatial	Low	Texture uncertainty
temporal_diff	Temporal	Low	Inter-frame motion energy
temporal_std	Temporal	Low	Temporal fluctuation
blockiness	Artifact	Low	Compression blocking severity

D. Models, Validation, and Runtime Protocol

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 2 August 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 447-463
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

Seven predictive models were evaluated. The proposed deployment target was a compact gradient-boosted decision tree (GBDT) with 50 trees and roughly 350 nodes. A wider GBDT and a random forest were included as stronger non-linear baselines; a decision tree served as a minimalist baseline. Two linear baselines measured how much signal comes from bitrate alone or from bitrate plus scale and codec metadata. Finally, a tiny multilayer perceptron tested whether a small neural model justified its extra implementation complexity on this feature space.

Validation used GroupKFold with five folds and shot identity as the grouping variable, so no shot appeared in both train and test partitions. This guard against content leakage is important because frame-derived statistics from adjacent frames or from the same shot are otherwise highly correlated. The primary metrics were mean absolute error (MAE), root mean squared error (RMSE), and Pearson correlation. MAE was emphasized because guardrail policies are usually discussed in absolute VMAF points.

Runtime was benchmarked in a single CPU-only environment. The paper therefore treats absolute latency as machine-specific and uses it mainly to compare relative deployment cost across alternatives. Predictor-only latency and full-chain latency are reported separately because some production stacks already possess decode and feature-extraction stages for adjacent telemetry tasks. The comparative analysis of various model configurations, including their serialized sizes and predictor-only latencies, is presented in Table 5.

Table 5. Model Configurations, Serialized Sizes, and Predictor-Only Latency

Model	Features	Configuration	Model size (KB)	Predictor latency (ms/sample)
Bitrate-only LR	1	Linear regression on log bitrate	0.25	0.017
Metadata LR	3	Linear regression on bitrate, scale, codec	0.84	0.031
Decision Tree	14	Depth 8, leaf size 2	5.42	0.081
Student GBDT	14	50 trees, ~350 nodes	43.44	0.484
Tiny MLP	14	2 hidden layers (24, 12)	97.91	0.612
Reference GBDT	14	180 trees, wider ensemble	197.67	1.175
Random Forest	14	400 trees	1785.76	15.520

E. Guard Thresholds and Tokenization

The proxy was also evaluated as an online guard at thresholds of 60, 70, and 80 VMAF. These cut points correspond to increasingly strict policies: reject clearly poor quality, gate borderline quality, and protect a higher quality target. Precision, recall, and F1 were computed on the same shot-level out-of-fold predictions used for regression scoring. To support AI-assisted operations, each shot prediction was serialized into a compact token containing shot identifier, codec, resolution, predicted VMAF, guard band, and a heuristic issue label. The issue label was assigned by deterministic rules over standardized features: high blockiness implied blockiness; low spatial-

detail measures under aggressive scaling implied detail loss; elevated motion measures combined with low predicted quality implied motion stress; otherwise the shot was labeled none. These labels are explanatory heuristics, not ground-truth artifact annotations.

IV. RESULT

A. Teacher Landscape and Codec Effects

The teacher labels spanned the intended operational range. Across the twelve variants, mean VMAF rose from 29.92 for H.264 at 180p-low to 98.20 for H.265 at 360p-high. This range is useful because it prevents the student from solving only a narrow, near-ceiling problem. At the same time, the corpus is small enough that the results should be read as a feasibility study rather than as a final benchmark. Within this excerpt and ladder, H.265 consistently improved quality at lower bitrate. The duration-weighted pairwise comparison shows an average bitrate saving of 15.23% relative to H.264 together with a mean VMAF gain of 5.71 points. The gain is positive in every matched pair, although its magnitude varies by operating point. Because this result comes from one animated excerpt, it should be interpreted as evidence that codec identity matters to the student, not as a universal estimate of H.265 efficiency.

Table 6. Duration-Weighted Matched Codec-Pair Comparison

Matched Pair	H.264 bitrate (kb/s)	H.265 bitrate (kb/s)	Bitrate saving (%)	Weighted ΔVMAF
180p-Low	179	152	15.08	5.51
180p-High	312	263	15.71	5.94
240p-Low	342	291	14.91	7.49
240p-High	608	519	14.64	5.04
360p-Low	986	832	15.62	2.87
360p-High	1478	1251	15.36	7.41
Average	—	—	15.23	5.71

Note. Table 6 uses duration-weighted pairwise means. Small differences relative to Table 3 arise from weighting and rounding.

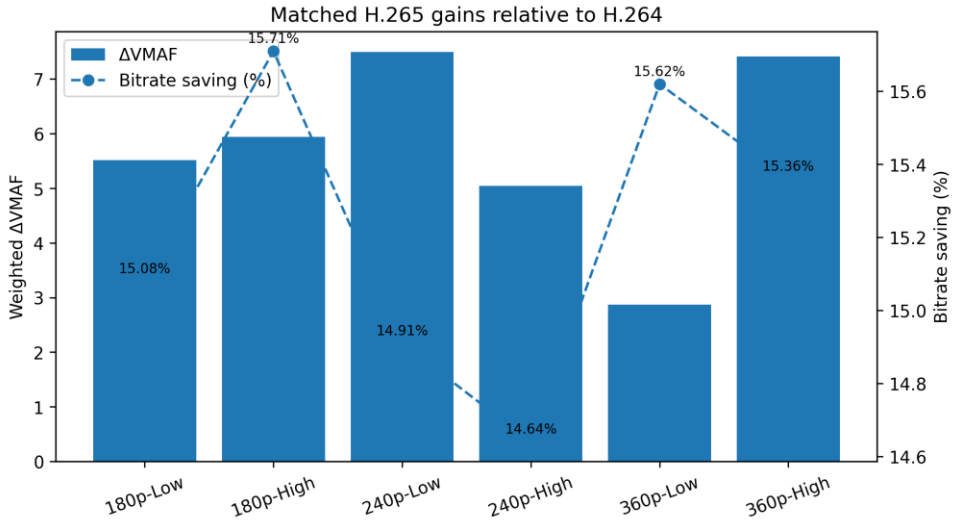

Figure 4. H.265 Gains Relative to Matched H.264 Operating Points; Bars Show Weighted ΔVMAF and the Dashed Line Shows Bitrate Saving

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 2 August 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 447-463
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

The near-ceiling values in the 360p stratum deserve careful interpretation. They reflect this specific animated content, the chosen ladder, and the standard full-reference practice of comparing upscaled distorted streams to the reference. They should not be overgeneralized into a claim that 360p is generally visually transparent. The performance gains and bitrate savings achieved by H.265 over H.264 across these strata are summarized in Table 6 and further illustrated in Figure 4.

B. Regression Performance

Table 6 reports the main regression results. On this corpus, the random forest achieved the lowest MAE at 6.37, followed closely by the wider GBDT at 6.46. The proposed student GBDT reached MAE = 6.56, RMSE = 8.32, and Pearson r = 0.913. The absolute gap between the student and the best-performing baseline was only 0.19 MAE points, whereas the student remained dramatically smaller and faster than the heavier ensembles.

Table 7. Overall Out-of-Fold Performance Across Model Families

Model	MAE	RMSE	Pearson r
Random Forest	6.37	8.00	0.920
Reference GBDT	6.46	8.14	0.917
Student GBDT	6.56	8.32	0.913
Metadata LR	7.35	9.02	0.896
Decision Tree	7.70	9.89	0.872
Bitrate-only LR	8.36	10.14	0.851
Tiny MLP	9.44	12.38	0.807

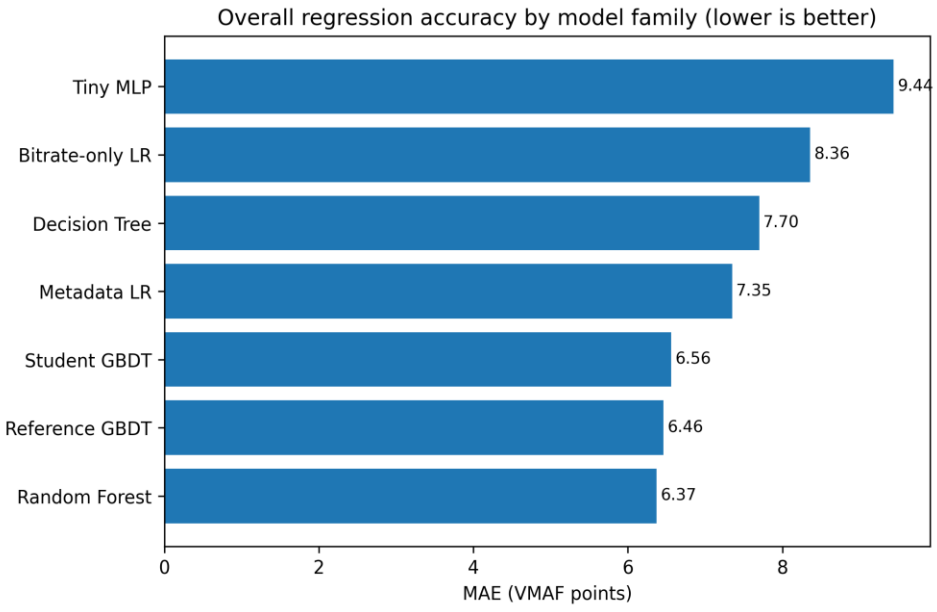

Figure 5. Overall Model MAE on the Shot-Grouped Evaluation. Lower is Better

Relative to the simplest regressors, the compact student delivered a meaningful improvement. Based on the reported rounded values, MAE fell by approximately 10.7% versus metadata-only

regression and 21.5% versus bitrate-only regression. The tiny MLP did not justify its extra complexity on this feature set, which supports the choice of a compact tree ensemble as the default deployment artifact. No claim of statistical significance is made here. With only 13 content groups, fold-level variation can materially affect the exact ranking of closely performing models. The more defensible conclusion is operational: among the tested models, the student GBDT offered the most attractive accuracy-cost balance. The performance metrics for all evaluated model families are documented in Table 7, with their respective Mean Absolute Error (MAE) scores compared in Figure 5.

C. Ablation and Stratified Behavior

The ablation study clarifies what the student actually learned. Removing temporal features raised MAE from 6.56 to 7.01, while removing spatial-detail features raised it further to 7.22. Dropping blockiness or the codec flag also degraded performance. This pattern supports a multi-cue interpretation of the student: bitrate and scale matter, but motion, texture retention, blocking, and codec-aware interactions also contribute materially to teacher alignment.

Table 8. Student Ablation Analysis

Ablation setting	MAE	RMSE
Student GBDT (full 14 features)	6.56	8.32
Without temporal features	7.01	8.81
Without spatial-detail features	7.22	9.03
Without blockiness	6.88	8.61
Without codec flag	6.91	8.76
Metadata only	7.35	9.02
Bitrate only	8.36	10.14

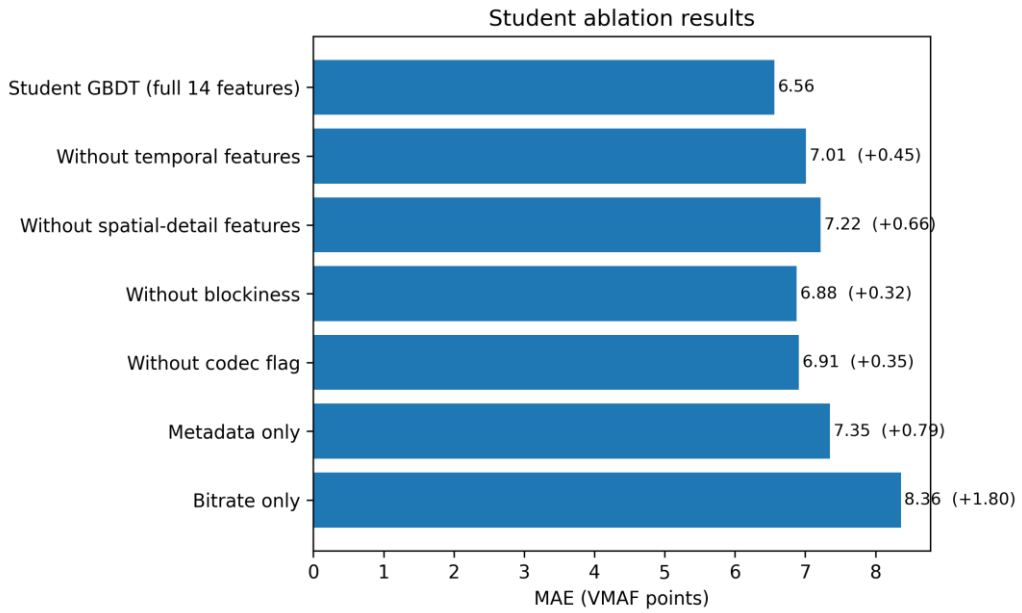

Figure 6. Ablation Results for the Student Proxy. Parenthetical Values Indicate MAE Increase Relative to the Full 14-Feature Model

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 2 August 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 447-463
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

Stratified results tell a similar story. The student was slightly more accurate on H.265 than on H.264, and its error decreased as quality approached the ceiling in the 360p stratum. That trend is unsurprising, because the most heavily distorted rungs produce wider perceptual variation. The important point is that the student remained useful across the entire ladder rather than only at the top end. The contribution of individual feature families to this performance is analyzed via ablation in Table 8 and Figure 6, while the detailed accuracy metrics stratified by codec and resolution are provided in Table 9 and Table 10, respectively.

Table 9. Student Accuracy Stratified by Codec

Codec stratum	Mean teacher VMAF	Student MAE	Student RMSE
H.264/AVC	63.41	6.86	8.63
H.265/HEVC	68.55	6.25	7.98

D. Guarding and Deployment Trade-Offs

The thresholded-guard results show that the proxy is useful for decision support, not merely for descriptive scoring. F1 ranged from 0.826 at 60 VMAF to 0.900 at 80 VMAF. The 70-VMAF threshold was a particularly balanced operating point, combining precision = 0.910 with F1 = 0.893. For practical release engineering, this is often more important than squeezing a few tenths of a point from regression error.

Table 10. Student Accuracy Stratified by Delivery Resolution

Resolution stratum	Mean teacher VMAF	Student MAE	Student RMSE
180p	43.22	7.21	9.12
240p	60.55	6.39	8.05
360p	94.16	5.92	7.74

Deployment measurements make the trade-off explicit. Predictor-only latency for the student GBDT was 0.484 ms per sample and the serialized model size was 43.44 KB. In the single-machine CPU benchmark, the full decode-feature-predict chain averaged 42.61 ms compared with 1117.41 ms for the teacher, a 26.22× end-to-end speed-up. The bitrate-only and metadata-only baselines were cheaper still, but their error was materially worse. The random forest was slightly more accurate yet much heavier in both model size and runtime.

Table 11. Student GBDT Quality-Guard Performance

Guard threshold	Precision	Recall	F1
60	0.818	0.833	0.826
70	0.910	0.877	0.893
80	0.873	0.928	0.900

The token interface completed the intended workflow. A shot prediction could be converted into a compact record that an LLM (Sun et al., 2024) can summarize without re-running signal processing. That is the right division of labor for this paper: the proxy performs deterministic estimation, while the language model performs explanation. The present study does not evaluate the downstream LLM (Zhao et al., 2024) quantitatively; it validates only that the interface is

compact and semantically legible. he classification performance of the quality-guard is detailed in Table 11 and Figure 7, while the comparative deployment efficiency is analyzed in Table 12 and Figure 8; finally, illustrative examples of the resulting LLM-ready tokens are presented in Table 13.

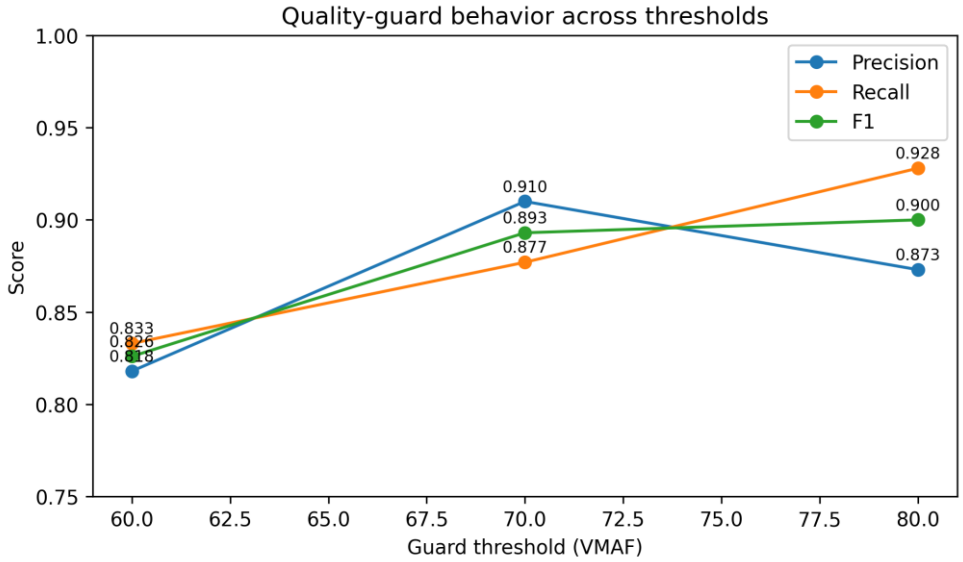

Figure 7. Precision, Recall, and F1 Across the Three Guard Thresholds

V. DISCUSSION

Several limitations materially constrain the scope of the findings. First, the corpus contains only one animated source excerpt. Live action, camera noise, film grain, screen content, and device-captured video could change both teacher behavior and feature usefulness. Second, the experiment uses only 13 shot groups. That is enough for a disciplined pilot with grouped validation, but it is too small for strong claims about statistical stability across content classes.

Table 12. Deployment Latency and Footprint Comparison

System	Latency (ms/sample)	Model size (KB)	Speed-up vs teacher
Teacher VMAF	1117.41	—	1.00×
Bitrate-only proxy	31.82	0.25	35.12×
Metadata proxy	34.60	0.84	32.30×
Student GBDT (predictor only)	0.484	43.44	2308.70×
Student GBDT full chain (decode + feature + predict)	42.61	43.44	26.22×
Reference GBDT full chain	43.31	197.67	25.80×
Random Forest full chain	57.62	1785.76	19.39×

Third, the distortion space is intentionally narrow: two codecs, three delivery resolutions, and two quality levels per pair. The paper therefore says little about AV1, VP9, rate-control modes beyond the chosen CRFs, rebuffering effects, display-adaptive models, or enhancement filters. Fourth,

Journal of Technology Informatics and Engineering (JTIE)
Vol. 4 No. 2 August 2025
E-ISSN: 2961-9068; P-ISSN: 2961-8215, Pages 447-463
Journal Link: https://jtie.stekom.ac.id/index.php/jtie/index

the latency numbers come from one CPU-only benchmark environment, so they should be interpreted as relative comparisons rather than universal serving times.

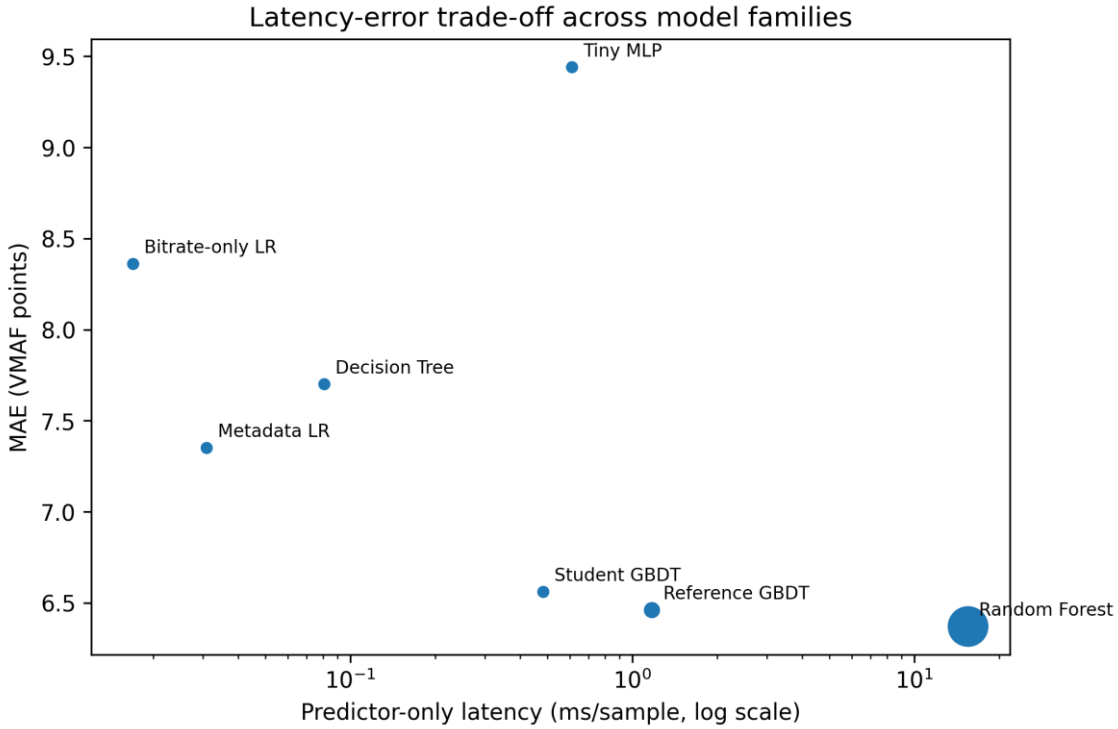

Figure 8. Predictor-Only Latency Versus MAE Across Model Families; Marker Area is Proportional to Serialized Model Size

Table 13. Example LLM-Ready Quality Tokens Produced by the Proxy

Shot	LLM-ready quality token	Natural-Language Incident Summary Seed
S03	{ "codec": "H264", "res": "240p", "pred_vmaf": 47.6, "band": "red", "issue": "blockiness" }	High-risk regression; visible blocking in textured motion.
S08	{ "codec": "H265", "res": "240p", "pred_vmaf": 71.8, "band": "amber", "issue": "detail_loss" }	Borderline quality; acceptable but below release target.
S12	{ "codec": "H265", "res": "360p", "pred_vmaf": 97.9, "band": "green", "issue": "none" }	No guard action; summarize as near-reference quality.

Note. The issue field is heuristic and intended for explanation support rather than for artifact-classification benchmarking.

Fifth, the LLM-ready issue labels are heuristic. They improve interpretability, but they were not validated against human artifact labels. Finally, the study evaluates VMAF distillation rather than subjective quality directly; it inherits any limitations of the teacher and does not replace the need for periodic human-facing validation.

VI. CONCLUSION AND RECOMMENDATION

On this pilot corpus, a compact gradient-boosted student preserved enough of VMAF's behavior to support useful edge-side quality guarding. The model achieved MAE = 6.56 and RMSE = 8.32 while remaining close to much heavier ensembles and substantially improving on simpler regressors. As a guard, it maintained strong threshold behavior, and as a deployment artifact it was small enough and fast enough to be practical in CPU-constrained environments.

The correct operational recommendation is therefore a two-tier architecture. Use the teacher offline for corpus refresh, ladder redesign, and periodic recalibration. Use the student online for continuous monitoring, fast regression detection, and low-cost quality summarization. Downstream LLMs should consume the compact tokens produced by the proxy rather than trying to infer perceptual quality from raw logs alone.

Future work should extend the corpus across content classes, codecs, devices, and viewing models, and should evaluate the token interface in end-to-end reporting workflows. The current evidence supports feasibility. It does not yet justify broad claims of production-wide generalization.

CrediT Author

B.Z. contributed to the conceptualization of the study, methodology development, experimental design, data collection, implementation, and initial manuscript drafting. H.W. contributed to the conceptualization of the study, supervision, methodology validation, formal analysis, manuscript review, and final approval of the submitted version. X.C. contributed to data processing, visualization, literature review, result interpretation, and manuscript revision. All authors have read and approved the final version of the manuscript. B.Z. and H.W. contributed equally to this work and should be regarded as co-first authors.

REFERENCES

Aaron, A., Li, Z., Manohara, M., Lin, J. Y., Wu, E. C.-H., & Kuo, C.-C. J. (2015). Challenges in Cloud Based Ingest and Encoding for High Quality Streaming Media. 2015 IEEE International Conference on Image Processing (ICIP), 1732–1736. https://doi.org/10.1109/icip.2015.7351101

Aistov, K., & Koroteev, M. (2023). VMAF Re-Implementation on PyTorch: Some Experimental Results. *arXiv*, 2310.15578. https://arxiv.org/abs/2310.15578

Bampis, C. G., Li, Z., & Bovik, A. C. (2019). Spatiotemporal Feature Integration and Model Fusion for Full-Reference Video Quality Assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, *29*(8), 2256–2270. https://doi.org/10.1109/tcsvt.2018.2861234

Bampis, C. G., Li, Z., Katsavounidis, I., & Bovik, A. C. (2018). Recurrent and Dynamic Models for Predicting Streaming Video Quality of Experience. *IEEE Transactions on Image Processing*, *27*(7), 3316–3331. https://doi.org/10.1109/tip.2018.2815842

Benjamin, N., Yulianingsih, S., & Marie, I. (2026). Explainable AI-Driven Strategic Decision-Making in SMEs: Simulation-Based Evaluation of Ethical Governance. *Journal of Management and Informatics*, *5*(1), 1–12. https://doi.org/10.51903/jmi.v3i1.314

Blender Foundation. (2008). *Big Buck Bunny* [Film]. https://peach.blender.org/

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., ... Amodei, D. (2020). Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems (NeurIPS)*, *33*, 1877–1901. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-paper.pdf

Chen, J., Sun, X., Wu, Q., & Jackson, M. (2024). Risk-Calibrated Biomedical Search: Calibrated Selection of LLM-Style Query Expansions on BEIR TREC-COVID. *Journal of Advanced Computing Systems*, *4*(4), 61–79. https://doi.org/10.69987/jacs.2024.40406

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT 2019*, 4171–4186. https://doi.org/10.18653/v1/n19-1423

Han, S., Mao, H., & Dally, W. J. (2016). Deep Compression: Compressing Deep Neural Networks with Pruning, Trained Quantization and Huffman Coding. *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1510.00149

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the Knowledge in a Neural Network. *arXiv*, 1503.02531. https://arxiv.org/abs/1503.02531

Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., Andreetto, M., & Adam, H. (2017). MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv*, 1704.04861. https://arxiv.org/abs/1704.04861

Huynh-Thu, Q., & Ghanbari, M. (2008). Scope of Validity of PSNR in Image/Video Quality Assessment. *Electronics Letters*, *44*(13), 800–801. https://doi.org/10.1049/el:20080522

ITU-R. (2019). *Methodologies for the Subjective Assessment of the Quality of Television Images (Recommendation ITU-R BT.500-14)*. https://www.itu.int/rec/r-rec-bt.500-14-201910-i/en

ITU-T. (2023). *Subjective Video Quality Assessment Methods for Multimedia Applications (Recommendation ITU-T P.910)*. https://www.itu.int/rec/t-rec-p.910-202307-i/en

Li, S., Zhang, F., Ma, L., & Ngan, K. N. (2011). Image Quality Assessment by Separately Evaluating Detail Losses and Additive Impairments. *IEEE Transactions on Multimedia*, *13*(5), 935–949. https://doi.org/10.1109/tmm.2011.2152382

Li, Y. (2024). Test-in-the-Loop LLM Repair: Verifiable Automated Program Repair on QuixBugs with a "Failing Test → Patch → Regression Test" Loop. *Journal of Advanced Computing Systems*, *4*(2), 62–75. https://doi.org/10.69987/jacs.2024.40206

Li, Z., Aaron, A., Katsavounidis, I., Moorthy, A. K., & Manohara, M. (2016). Toward a Practical Perceptual Video Quality Metric. *Netflix Tech Blog*. https://netflixtechblog.com/toward-a-practical-perceptual-video-quality-metric-653f207b4e20

Malolo, A. M. I. H. B., Sampetoding, E. A. M., Octavian, O., & Gomantara, J. (2025). Analysis of the Use of Interactive Video Features on the Cookpad Application for Culinary MSMEs Using TAM and SUS. *Jurnal Ilmiah Sistem Informasi*, *4*(3), 932–943. https://doi.org/10.51903/k7ta4t87

Mittal, A., Moorthy, A. K., & Bovik, A. C. (2012). No-Reference Image Quality Assessment in the Spatial Domain. *IEEE Transactions on Image Processing*, *21*(12), 4695–4708. https://doi.org/10.1109/tip.2012.2214050

Mittal, A., Soundararajan, R., & Bovik, A. C. (2013). Making a Completely Blind Image Quality Analyzer. *IEEE Signal Processing Letters*, *20*(3), 209–212. https://doi.org/10.1109/lspl.2012.2227726

Netflix. (2025a). *VMAF: Video Multi-Method Assessment Fusion* [Software]. https://github.com/netflix/vmaf

Netflix. (2025b). Models. *the VMAF GitHub Repository*. https://github.com/netflix/vmaf/blob/master/resource/doc/models.md

Netflix. (2025c). Using VMAF with FFmpeg. *the VMAF GitHub Repository*. https://github.com/netflix/vmaf/blob/master/resource/doc/ffmpeg.md

Rassool, R. (2017). VMAF Reproducibility: Validating a Perceptual Practical Video Quality Metric. *2017 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)*, 1–2. https://doi.org/10.1109/bmsb.2017.7986164

Sheikh, H. R., & Bovik, A. C. (2006). Image Information and Visual Quality. *IEEE Transactions on Image Processing*, *15*(2), 430–444. https://doi.org/10.1109/tip.2005.859378

Sun, X., Chen, J., Zhou, B., & Kuo, M.-J. (2024). ConRAG: Contradiction-Aware Retrieval-Augmented Generation under Multi-Source Conflicting Evidence. *Journal of Advanced Computing Systems*, *4*(7), 50–64. https://doi.org/10.69987/jacs.2024.40705

Tan, B. L., Liem, C. A., & Amen, M. (2026). Efficient Temporal Segmentation and Classification of Short-Form Video Content Using Lightweight CNN-LSTM Architecture. *Journal of Technology Informatics and Engineering*, *5*(1), 1–16. https://doi.org/10.51903/jtie.v5i1.441

Tan, M., & Le, Q. (2019). EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. *Proceedings of ICML 2019*, 6105–6114. https://arxiv.org/abs/1905.11946

Wang, Z., Bovik, A. C., Sheikh, H. R., & Simoncelli, E. P. (2004). Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, *13*(4), 600–612. https://doi.org/10.1109/tip.2003.819861

Wang, Z., Simoncelli, E. P., & Bovik, A. C. (2003). Multiscale Structural Similarity for Image Quality Assessment. *The 37th Asilomar Conference on Signals, Systems & Computers*, *2*, 1398–1402. https://doi.org/10.1109/acssc.2003.1292216

Zhao, S., Wang, H., & Davison, N. (2024). Profit-Maximizing Cost-Sensitive Credit Scoring with LLM-Extracted Policy Constraints. *Journal of Advanced Computing Systems*, *4*(3), 91–108. https://doi.org/10.69987/jacs.2024.40307