Computation and Performance Studies
Volume 1 · Number 1 · 2026

A Quality-Latency Frontier Changes with the Platform: Joining Published Image-Assessment Tables

Dominic Knight[1] · Tyler Robertson[2] · Evan Anderson[3]

[1] Australia · [2] New Zealand · [3] Netherlands

Original publication: October 28, 2026 | Article 616

Abstract

Choosing an image-quality metric requires more than ordering its reported correlations. An exact-name join of five quality panels and two timing columns retains eleven complete methods from a pinned IQA-PyTorch snapshot. Five methods are nondominated for CPU latency and average reported SRCC, while only two remain on the corresponding GPU frontier. The quality-log-latency slopes are uncertain on both platforms. These conditional frontiers describe a published table join, with different quality and timing workloads, rather than a new controlled hardware benchmark or a universal recommendation.

A Quality-Latency Frontier Changes with the Platform: Joining Published Image-Assessment Tables

Abstract

Choosing an image-quality metric requires more than ordering its reported correlations. An exact-name join of five quality panels and two timing columns retains eleven complete methods from a pinned IQA-PyTorch snapshot. Five methods are nondominated for CPU latency and average reported SRCC, while only two remain on the corresponding GPU frontier. The quality-log-latency slopes are uncertain on both platforms. These conditional frontiers describe a published table join, with different quality and timing workloads, rather than a new controlled hardware benchmark or a universal recommendation.

1. A fast metric is defined relative to a quality objective

Image assessment is often treated as a ranking problem in which the method with the highest correlation wins. A system choosing a metric for repeated use also faces a computational constraint. A slightly stronger reported correlation may or may not justify a much longer call, and the answer can depend on the platform used to execute it. The research question is which reported methods remain nondominated when a common quality summary is considered jointly with CPU latency and then with GPU latency.

The semantic and distortion distinction proposed in Meng, 2026 motivates attention to what a quality metric measures. It does not determine which computational burden a workflow can accept. We do not run that model or compare its measured speed with existing methods. Our evidence is a pinned public benchmark snapshot from the IQA-PyTorch resource, referenced as Chen and Mo, 2022. It contains separate published quality and efficiency tables that can be joined transparently without pretending that a new experiment has been performed.

The resulting frontier is conditional on several explicit choices. Quality is the arithmetic mean of reported absolute SRCC across five panels. Latency is one of two source columns. Missing records are excluded, metric identifiers must match exactly, and dominance requires no worse values on both objectives with a strict improvement on at least one. A frontier can change when any of those definitions changes. Its purpose is to expose the tradeoff available in the captured records, rather than to hide those choices behind the claim that one metric is best overall.

2. Measurement and workload before interpretation

The computational perspective in Williams et al., 2009 makes platform and workload central to performance interpretation. The benchmarking concerns in Mytkowicz et al., 2009 and the measurement approach in Kalibera and Jones, 2013 further caution against treating a single reported time as a universal constant. We therefore preserve the published timing values as source summaries. We do not attach invented repeated-run standard errors, infer energy use, or equate a reported wall-clock call time with training cost. The source table does not provide the information needed to estimate those quantities.

The reported efficiency helper times path-based metric calls, with source documentation describing average evaluation at 1080 by 800 resolution. Model construction is outside the timed region, and the first image timing is omitted. A current helper also invokes CUDA synchronization, including in its CPU branch, leaving limits on reconstructing all historic table cells from that helper alone. These details matter because our join is a comparison of reported columns, not a claim that a newly executed script reproduced the timing conditions or controlled every platform difference.

The quality side uses five panels: LIVE Challenge, KonIQ-10k, TID2013, FLIVE, and SPAQ. The selected resources include smartphone photographs (Fang et al., 2020), diverse natural images (Hosu et al., 2019), and controlled artificial distortions (Ponomarenko et al., 2015). Other assessment settings include aesthetic judgments (Murray et al., 2012) and perceptual restoration assessment (Gu et al., 2020); those two resources are contextual references and are not panels in this calculation. The five-panel arithmetic mean assigns equal panel weight and does not define a common latent perceptual quantity. It is neither image-count weighted nor a claim about the utility of every application using a metric.

Metric design also varies. Mittal et al., 2012 and Mittal et al., 2013 use visual-statistics approaches, while Talebi and Milanfar, 2017 learns assessment outputs. Yang et al., 2022 uses multidimensional attention and Ying et al., 2019 addresses picture-quality assessment. Their inclusion in a common table enables a resource comparison but does not erase differences in training data, intended targets, or implementation. Those differences can affect both the benchmark correlations and the observed call time. We do not attribute a frontier position to a single architectural feature or infer a causal benefit of attention, scale, or training size.

3. Constructing the exact comparison cohort

We read the pinned no-reference quality table and retain methods with nonmissing entries in its first five panels: LIVE Challenge, KonIQ-10k, TID2013, FLIVE, and SPAQ. Each cell stores PLCC, SRCC, and KRCC in a slash-separated format. We extract the SRCC component as reported. The source uses correlation magnitudes; the mean therefore summarizes reported absolute rank association rather than signed association on a common raw score direction. The sparsely reported KonIQ-1024 column is excluded as a separate resolution condition.

Twelve method rows have all five selected quality values. We join them by exact metric identifier to the efficiency table and require both CPU Xe6150 and CUDA V100 timings. Eleven remain after the join. No spelling alias is guessed and no missing latency is imputed. The final cohort contains brisque, nrqm, niqe, ilniqe, nima, pi, cnniqa, paq2piq, dbcnn, topiq_nr, and maniqa. This explicit list defines the comparison population and prevents a partially reported method from being treated as dominated merely because its timing is absent.

The mean quality score assigns equal weight to the five panels. It is calculated once and used for both platform frontiers. Latencies must be positive so that their logarithms can be calculated for the descriptive regressions. Family labels group related metric names for the covariance calculation; the eleven records form ten such groups. Exact-ID joining and family grouping serve different purposes. The former determines which numeric records are compared, while the latter expresses a working dependence structure among those retained methods.

For a given platform, method A dominates method B when A has latency no greater than B and mean quality no lower than B, with at least one strict inequality. Every method not dominated by any retained method is on the platform-specific frontier. This is a finite-table relation and does not require a regression model. Equal values are treated as ties at the published precision. No tolerance is added to imply knowledge of an unreported numerical difference, and no missing method is silently assigned a poor score.

4. Research model and descriptive regression

Figure 1 shows the two sources entering an exact-ID join before platform-specific comparisons. The quality summary is held constant across platforms, so changes in frontier membership arise from the reported latency ordering within the retained cohort. This does not make the platform a randomized intervention. The CPU and GPU columns can differ in hardware, execution behavior, and measurement context. They supply alternative operational records for the same method identifiers, not an isolated causal estimate of moving an arbitrary workload to a GPU.

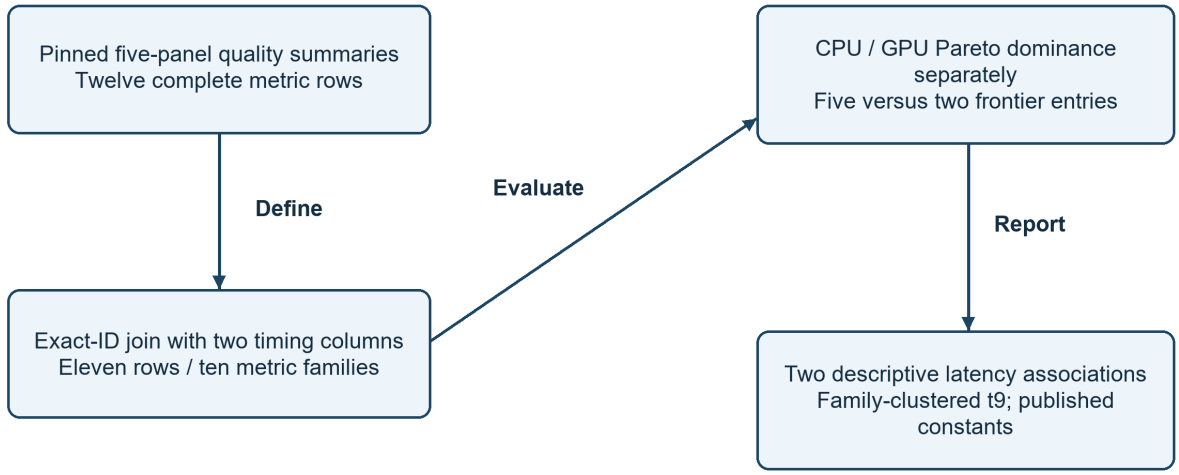


Figure 1. *Published quality summaries and timing records are joined by exact method identifier. CPU and GPU dominance are assessed separately, with a common quality aggregation and a complete-case cohort.*

Two OLS models relate mean reported SRCC to log latency, one using the CPU column and the other the GPU column. These associations describe whether slower records tend to have higher quality summaries in the retained table. They do not determine the frontier, estimate an efficiency production function, or establish an intrinsic speed-quality law. Carroll et al., 2006 makes measurement uncertainty relevant to such interpretation, but the source lacks image-level score uncertainty and repeated timing records needed to propagate it into a more comprehensive model.

Covariance is clustered by ten declared method families, with finite-sample correction and a t reference of nine degrees of freedom. Cameron and Miller, 2015 explains why related implementations should not automatically be treated as independent observations. White, 1980 supplies the broader heteroskedasticity-aware foundation. Family grouping still omits shared benchmark panels, training ancestry, and source measurement conditions. The intervals are working descriptions of a small captured cohort, conditional on the rounded published numbers, rather than confidence statements about every possible image-quality algorithm.

Table 1. *Mean reported absolute SRCC regressed on log published latency separately for CPU and GPU. N is eleven methods. Both models include an intercept, use ten family clusters with finite-sample correction, and use a t reference with nine degrees of freedom. There are no additional controls.*

Model	Term	Estimate (SE)	95% CI	N
M1	Log reported CPU seconds	0.015323 (0.021419)	[-0.033130, 0.063776]	11
M1: fit	R-squared = 0.0510	-	-	11
M2	Log reported GPU seconds	-0.001230 (0.034870)	[-0.080110, 0.077651]	11
M2: fit	R-squared = 0.0001	-	-	11

Model labels: M1 = Mean reported SRCC versus log CPU latency; M2 = Mean reported SRCC versus log GPU latency.

5. Frontier membership and uncertain average slopes

The CPU frontier contains brisque, nima, cnniqa, paq2piq, and topiq_nr. The GPU frontier contains paq2piq and topiq_nr. Thus three CPU-nondominated choices become dominated under the alternative timing column, while the two GPU frontier entries are also CPU frontier entries. Six of the eleven methods are dominated under both platform records. These statements are exact for the selected objectives and published precision. They do not mean that the six methods have no useful property absent from the quality and latency summaries.

The CPU log-latency coefficient is positive 0.015323, with standard error 0.021419 and a 95% interval from negative 0.033130 to positive 0.063776. The model R-squared is 0.0510. The GPU coefficient is negative 0.001230, with standard error 0.034870, an interval from negative 0.080110 to positive 0.077651, and R-squared approximately 0.0001. Neither association shows a clear average slope under the stated inference. A near-zero fitted linear slope also does not make every method an equally good resource choice.

That last distinction is central. Dominance is pairwise and permits a nonlinear, irregular frontier, whereas the regression summarizes one global linear association with log time. A table can contain clearly dominated entries even when its overall slope is uncertain. Conversely, a positive slope would not identify which specific method dominates another. Reporting the frontier alongside the model keeps the decision relation and the descriptive trend separate, without using the inferential table as a substitute for checking the actual retained records.

Rank-correlation traditions in Spearman, 1904 and Kendall, 1938 explain the association quantities reported in the source panels, but their presence does not justify averaging them as if all panels measured one latent construct. Demšar, 2006 adds a relevant caution about comparisons over related benchmark tasks. The five panels are not five independent replications of a controlled hardware tradeoff. They provide one chosen quality summary, and a workflow emphasizing a single dataset could legitimately obtain a different frontier from the same source methods.

6. Choosing among nondominated records requires more evidence

The frontier deliberately omits deployment utility. Gneiting and Raftery, 2007 motivates stating what a predictive evaluation rewards, and Willmott and Matsuura, 2005 clarifies that a numerical error summary depends on its objective. Here mean absolute SRCC rewards rank association averaged across the selected panels. It does not quantify the cost of a mistaken image decision, calibration of a quality threshold, or performance on a rare distortion. A method could be dominated in the present two-dimensional summary while remaining attractive under one of those unmeasured criteria.

Other visual evaluation targets illustrate that limitation. Wu et al., 2023 concerns preference-related generated-image assessment, while Zhang et al., 2018 supplies a perceptual distance between images. Neither is reducible to the exact five-panel no-reference average used here. We therefore do not use frontier status as a recommendation for human preference, semantic alignment, or image restoration fidelity. Likewise, a smaller reported call time does not establish lower peak system memory, energy, or end-to-end application latency once loading, batching, and data movement are included.

Numerical and diagnostic caution also matters. The measurement discussion in Stevens, 1946 calls for identifying the meaning of every axis, and Cook, 1977 reminds us that a small regression can depend strongly on particular records. We did not delete influential metrics or tune the aggregation to strengthen a trend. The reporting perspective in Wasserstein and Lazar, 2016 favors retaining the uncertain slopes and conditional membership lists instead of turning either into a binary significance story. Rounding could affect a near-tie, and missing methods could change the frontier if complete records became available.

The snapshot, join cohort, parsing decisions, quality means, platform values, and family labels are retained according to documentation aims in Gebru et al., 2021 and reuse principles in Wilkinson et al., 2016. Those records allow a reader to change a panel weighting or a completeness rule explicitly. They do not establish that every historical timing cell can be regenerated from the current helper. A controlled follow-up would need common workloads, repeated timings, image-level quality outputs, and an application-specific objective before it could support a stronger method recommendation.

Completeness is part of the result, rather than a neutral preprocessing convenience. Requiring all five quality panels and both timings removes records that may be competitive under a smaller evidence requirement. We cannot infer their frontier positions without those missing entries. Conversely, using each method's available panels would compare different quality summaries and could reward selective reporting. The complete-case rule makes the current comparison coherent, but its cost is a narrower method cohort and an explicit inability to rank incompletely reported alternatives.

The frontier also supplies no preferred point among its surviving entries. Choosing between a faster lower-quality record and a slower higher-quality one requires a latency budget, a quality floor, or an application utility. None was elicited for this study. We consequently retain the full nondominated list for each platform and avoid a recommendation disguised as a geometric property. An institution can apply its own constraint to these captured summaries, provided it first checks whether the source workloads and dataset mixture resemble the intended use.

7. Conclusion

Joining the published tables yields five CPU-nondominated methods and two GPU-nondominated methods under the same quality summary. The average quality-log-latency associations remain uncertain, emphasizing that a frontier and a global regression answer different questions. The useful result is a transparent, platform-conditional comparison of complete reported records. Its practical interpretation depends on the selected panels, the timing workload, missing entries, and the decision objective that a real application needs to serve.

References

1. Yang, Sidi, et al. "MANIQA: Multi-dimension Attention Network for No-Reference Image Quality Assessment." *arXiv preprint arXiv:2204.08958*, 2022. https://arxiv.org/abs/2204.08958.
2. Fang, Yuming; Zhu, Hanwei; Zeng, Yan; Ma, Kede; Wang, Zhou. (2020). Perceptual Quality Assessment of Smartphone Photography. CVPR 2020, 3677-3686. https://openaccess.thecvf.com/content_CVPR_2020/html/Fang_Perceptual_Quality_Assessment_of_Smartphone_Photography_CVPR_2020_paper.html.
3. Cook, R. Dennis. (1977). Detection of Influential Observation in Linear Regression. Technometrics. https://doi.org/10.1080/00401706.1977.10489493.
4. White, Halbert. “A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity.” *Econometrica*, vol. 48, no. 4, 1980, pp. 817-838. https://doi.org/10.2307/1912934.
5. Mittal, A., et al. "Making a “Completely Blind” Image Quality Analyzer" *IEEE Signal Processing Letters*, 2013. https://doi.org/10.1109/lsp.2012.2227726.
6. Gu, Jinjin; Cai, Haoming; Chen, Haoyu; Ye, Xiaoxing; Ren, Jimmy; Dong, Chao. (2020). PIPAL: a Large-Scale Image Quality Assessment Dataset for Perceptual Image Restoration. https://arxiv.org/abs/2007.12142.
7. Chen, Chaofeng; Mo, Jiadi. (2022). IQA-PyTorch: PyTorch Toolbox for Image Quality Assessment. Software repository. https://github.com/chaofengc/IQA-PyTorch/tree/d47b7ac72ed25388371d0b637076c125cdb95a3c.
8. Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, IJsbrand Jan; Appleton, Gabrielle; Axton, Myles; Baak, Arie; Blomberg, Niklas; Boiten, Jan-Willem; da Silva Santos, Luiz Bonino; Bourne, Philip E.; Bouwman, Jildau; Brookes, Anthony J.; Clark, Tim; Crosas, Mercè; Dillo, Ingrid; Dumon, Olivier; Edmunds, Scott; Evelo, Chris T.; Finkers, Richard; Gonzalez-Beltran, Alejandra; Gray, Alasdair J.G.; Groth, Paul; Goble, Carole; Grethe, Jeffrey S.; Heringa, Jaap; ’t Hoen, Peter A.C; Hooft, Rob; Kuhn, Tobias; Kok, Ruben; Kok, Joost; Lusher, Scott J.; Martone, Maryann E.; Mons, Albert; Packer, Abel L.; Persson, Bengt; Rocca-Serra, Philippe; Roos, Marco; van Schaik, Rene; Sansone, Susanna-Assunta; Schultes, Erik; Sengstag, Thierry; Slater, Ted; Strawn, George; Swertz, Morris A.; Thompson, Mark; van der Lei, Johan; van Mulligen, Erik; Velterop, Jan; Waagmeester, Andra; Wittenburg, Peter; Wolstencroft, Katherine; Zhao, Jun; Mons, Barend. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. https://doi.org/10.1038/sdata.2016.18.
9. Talebi, Hossein and Peyman Milanfar. "NIMA: Neural Image Assessment" *arXiv preprint arXiv:1709.05424*, 2017. https://arxiv.org/abs/1709.05424.
10. Mittal, A.; Moorthy, A. K.; Bovik, A. C.. (2012). No-Reference Image Quality Assessment in the Spatial Domain. IEEE Transactions on Image Processing. https://doi.org/10.1109/tip.2012.2214050.
11. Wu, Xiaoshi, et al. "Human Preference Score v2: A Solid Benchmark for Evaluating Human Preferences of Text-to-Image Synthesis" *arXiv preprint arXiv:2306.09341*, 2023. https://arxiv.org/abs/2306.09341.
12. Murray, Naila; Marchesotti, Luca; Perronnin, Florent. (2012). AVA: A large-scale database for aesthetic visual analysis. 2012 IEEE Conference on Computer Vision and Pattern Recognition. https://doi.org/10.1109/cvpr.2012.6247954.
13. Hosu, Vlad, et al. "KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment" *arXiv preprint arXiv:1910.06180*, 2019. https://arxiv.org/abs/1910.06180.
14. Ponomarenko, Nikolay; Jin, Lina; Ieremeiev, Oleg; Lukin, Vladimir; Egiazarian, Karen; Astola, Jaakko; Vozel, Benoit; Chehdi, Kacem; Carli, Marco; Battisti, Federica; Jay Kuo, C.-C.. (2015). Image database TID2013: Peculiarities, results and perspectives. Signal Processing: Image Communication. https://doi.org/10.1016/j.image.2014.10.009.
15. Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. “Measurement Error in Nonlinear Models.” 2006. https://doi.org/10.1201/9781420010138.

16. Ying, Zhenqiang, et al. "From Patches to Pictures (PaQ-2-PiQ): Mapping the Perceptual Space of Picture Quality." *arXiv preprint arXiv:1912.10088*, 2019. https://arxiv.org/abs/1912.10088.

17. Cameron, A. Colin, and Douglas L. Miller. “A Practitioner’s Guide to Cluster-Robust Inference.” *Journal of Human Resources*, vol. 50, no. 2, 2015, pp. 317-372. https://doi.org/10.3368/jhr.50.2.317.

18. Demšar, Janez. (2006). Statistical Comparisons of Classifiers over Multiple Data Sets. Journal of Machine Learning Research. https://www.jmlr.org/papers/v7/demsar06a.html.

19. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.

20. Gneiting, Tilmann; Raftery, Adrian E. (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. Journal of the American Statistical Association. https://doi.org/10.1198/016214506000001437.

21. Zhang, Richard, et al. "The Unreasonable Effectiveness of Deep Features as a Perceptual Metric" *arXiv preprint arXiv:1801.03924*, 2018. https://arxiv.org/abs/1801.03924.

22. Stevens, S. S.. (1946). On the Theory of Scales of Measurement. Science. https://doi.org/10.1126/science.103.2684.677.

23. Meng, Zijie. (2026). Decoupling Semantics from Distortions: Multi-Scale Two-Stream Vision-Language Alignment for AI-Generated Image Quality Assessment. arXiv:2606.16799. https://arxiv.org/abs/2606.16799.

24. KENDALL, M. G.. (1938). A NEW MEASURE OF RANK CORRELATION. Biometrika. https://doi.org/10.1093/biomet/30.1-2.81.

25. Spearman, C.. (1904). The Proof and Measurement of Association between Two Things. The American Journal of Psychology. https://doi.org/10.2307/1412159.

26. Mytkowicz, Todd; Diwan, Amer; Hauswirth, Matthias; Sweeney, Peter F.. (2009). Producing wrong data without doing anything obviously wrong!. Proceedings of the 14th international conference on Architectural support for programming languages and operating systems. https://doi.org/10.1145/1508244.1508275.

27. Kalibera, Tomas; Jones, Richard. (2013). Rigorous benchmarking in reasonable time. Proceedings of the 2013 international symposium on memory management. https://doi.org/10.1145/2464157.2464160.

28. Wasserstein, Ronald L.; Lazar, Nicole A.. (2016). The ASA Statement on p -Values: Context, Process, and Purpose. The American Statistician. https://doi.org/10.1080/00031305.2016.1154108.

29. Willmott, CJ; Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. Climate Research. https://doi.org/10.3354/cr030079.

30. Williams, Samuel; Waterman, Andrew; Patterson, David. (2009). Roofline: an insightful visual performance model for multicore architectures. Communications of the ACM. https://doi.org/10.1145/1498765.1498785.