Benchmarking and Comparative Evaluation
Volume 1 · Number 1 · 2026

A Stable Winner Can Hide an Unstable Ranking: Correlation Criteria in Published Image-Quality Benchmarks

Dominic Johnston[1] · Daniel Nelson[2] · Gavin Bell[3]

[1] Denmark · [2] Norway · [3] Italy

Original publication: October 6, 2026 | Article 698

Abstract

Choosing a correlation coefficient can change a benchmark ordering even when its winner survives. We compared Pearson- and Spearman-based ranks in pinned IQA-PyTorch tables, retaining seventeen complete full-reference models and twelve complete no-reference models. Across nine dataset panels, only one leading model changed, while strict pairwise reversals ranged from two to thirty-four. Average normalized rank displacement was 0.114 across the twenty-nine model records. Missing entries also changed which models could compete. Benchmark claims should distinguish winner stability, ordering stability, and candidate-set coverage.

Exploratory audit of captured benchmark aggregates. Reported model correlations are reused as data; no quality network was trained or executed.

A Stable Winner Can Hide an Unstable Ranking: Correlation Criteria in Published Image-Quality Benchmarks

> Exploratory audit of captured benchmark aggregates. Reported model correlations are reused as data; no quality network was trained or executed.

Abstract

Choosing a correlation coefficient can change a benchmark ordering even when its winner survives. We compared Pearson- and Spearman-based ranks in pinned IQA-PyTorch tables, retaining seventeen complete full-reference models and twelve complete no-reference models. Across nine dataset panels, only one leading model changed, while strict pairwise reversals ranged from two to thirty-four. Average normalized rank displacement was 0.114 across the twenty-nine model records. Missing entries also changed which models could compete. Benchmark claims should distinguish winner stability, ordering stability, and candidate-set coverage.

1. What does a robust benchmark recommendation require?

A benchmark often communicates its findings through a sorted table. Readers may interpret a clear leading row as evidence that the comparison is stable, although the rest of the ordering can depend on the evaluation criterion. Image-quality assessment makes this problem concrete because the agreement between a computational score and human judgments can be summarized through linear association or rank association. Those summaries answer related questions without defining the same ordering of candidate systems.

Our research question is whether substituting Spearman correlation for Pearson correlation changes model ranks and selected leaders within a fixed, published quality-assessment table. The object of analysis is a captured report, rather than newly generated images or newly estimated correlations. This distinction permits a reproducible audit of what the report supports while preventing aggregate coefficients from being mistaken for access to image-level observations or independent experimental replications.

The separation between semantics and distortions proposed by Meng, 2026 makes evaluation alignment a relevant concern for generated-image assessment. A richer assessment architecture still requires a declared rule for judging agreement with human scores. The present evidence does not test that architecture or establish its relative accuracy. It examines a prior reporting decision: whether an apparent ranking remains recognizable after the numerical association criterion changes.

Three propositions organize the audit. First, retaining the same first-ranked model need not preserve the remaining pairwise order. Second, normalizing rank displacement can express criterion sensitivity without treating Pearson and Spearman coefficients as measurements on an interchangeable cardinal scale. Third, restricting a table to completely reported models can change its decision set. These are descriptive propositions about benchmark construction, not causal hypotheses about the benefits of reference images.

2. Association, agreement, and benchmark scope

Rank association has a long measurement history Spearman, 1904, and alternative rank summaries emphasize different features of concordance Kendall, 1938. Here we exchange only the two specified selection criteria; the published Kendall column is preserved in the acquired data but does not become a third opportunity to choose a favorable conclusion. The distinction between measurement scales Stevens, 1946 helps explain why raw correlation differences alone do not tell a reader how many candidate positions have changed.

Correlation is also not a general certificate of agreement. Bland and Altman, 1986 motivates care when interpreting association between measurement methods, while measurement-error analysis Carroll et al., 2006 warns that observed scores may combine several sources of variation. Neither reference licenses reconstructing uncertainty from a rounded aggregate coefficient without its underlying paired observations. Accordingly, this study never conducts a test that one reported image correlation significantly exceeds another.

Quality databases themselves cover different observations and tasks. KonIQ-10k concerns ecological image-quality assessment Hosu et al., 2019, whereas TID2013 Ponomarenko et al., 2015 and KADID-10k Lin et al., 2019 provide different distorted-image evaluation resources. Those sources identify benchmark context; their samples are not pooled here. Newer generated-image resources, including AGIQA-3K Li et al., 2023 and AIGIQA-20K Li et al., 2024, further illustrate why a table assembled for one setting cannot automatically certify performance in another.

The model menu is similarly heterogeneous. Deep perceptual distances Zhang et al., 2018 and structure-texture comparisons Ding et al., 2020 represent different scoring constructions. A multi-scale quality transformer Ke et al., 2021 introduces another modeling approach. Their presence in a common report does not make their training histories exchangeable. We therefore interpret a change in their relative position as a property of the captured comparison, without attributing that change to a particular architectural mechanism.

3. A pinned public report and a declared comparison set

We acquired the full-reference and no-reference benchmark CSV files from Chen and Mo, 2022, together with its README, evaluation script, correlation helper, and license text. The repository snapshot is pinned at commit d47b7ac72ed25388371d0b637076c125cdb95a3c. Local acquisition records retain exact file sizes and SHA256 values. The source script reports rounded absolute correlations, and its documentation specifies Pearson correlation without test-time value correction. Consequently, the audit concerns the released coefficient magnitudes and their stated protocol; discarded signs and unrounded values cannot be recovered.

The full-reference table contains seventeen model records and four complete dataset columns: CSIQ, LIVE, TID2008, and TID2013. All seventeen are retained. The no-reference table contains twenty-five model records. We use its first five dataset columns: LIVE Challenge, KonIQ-10k, TID2013, FLIVE, and SPAQ. A sixth column, KonIQ-10k at the reported 1024 setting, has extremely sparse coverage and is excluded from the primary design before computing the balanced comparison.

Twelve no-reference models have entries in all five selected columns. The thirteen others remain in the source file but are excluded from the balanced cohort. Nothing fills their missing cells, and exclusion does not imply weak performance. It defines a narrower comparison among completely reported candidates. The two primary cohorts therefore produce 128 model-dataset cells: sixty-eight full-reference cells and sixty no-reference cells, organized into nine separate table panels.

Some retained systems originate in statistical natural-image approaches Mittal et al., 2013 or spatial-domain quality assessment Mittal et al., 2012; others use feature similarity Lin Zhang et al., 2011 or neural assessment Talebi and Milanfar, 2017. These references identify different score families rather than supplying missing benchmark results. Multi-dimensional attention Yang et al., 2022 and language-image assessment Wang et al., 2022 also inform the landscape, although not every available variant survives the complete-record filter. We keep that coverage distinction visible instead of equating absence from the primary cohort with absence from quality research.

4. Rank displacement and conditional regression summaries

Within each panel, the largest reported coefficient receives rank one. Ties would receive average ranks, although none of the primary pair comparisons is tied at the captured precision. For each model, absolute displacement is the distance between its Pearson and Spearman positions divided by the number of competing models minus one. This normalization maps the maximum possible movement to one. It compares movement through a decision list, rather than claiming equal sampling precision across datasets.

A second outcome counts strict reversals among every unordered model pair. A pair reverses when the difference between its two coefficients has opposite signs under the two criteria. The denominator is the full set of pairs within that panel, and tied comparisons are counted separately. Leader identity provides a third, deliberately coarser summary. These three quantities can disagree legitimately because they describe list-wide movement, pairwise decisions, and the single top recommendation respectively.

Figure 1 shows the measurement sequence. Each model's normalized displacements are averaged over its four or five retained datasets. The resulting twenty-nine model averages form the regression dataset. Model 1 estimates their equal-model mean. Model 2 adds a no-reference-cohort indicator. Its coefficient compares two captured cohorts evaluated on different dataset menus; it cannot identify a causal effect of removing reference images or rank the overall quality of those methodological classes.

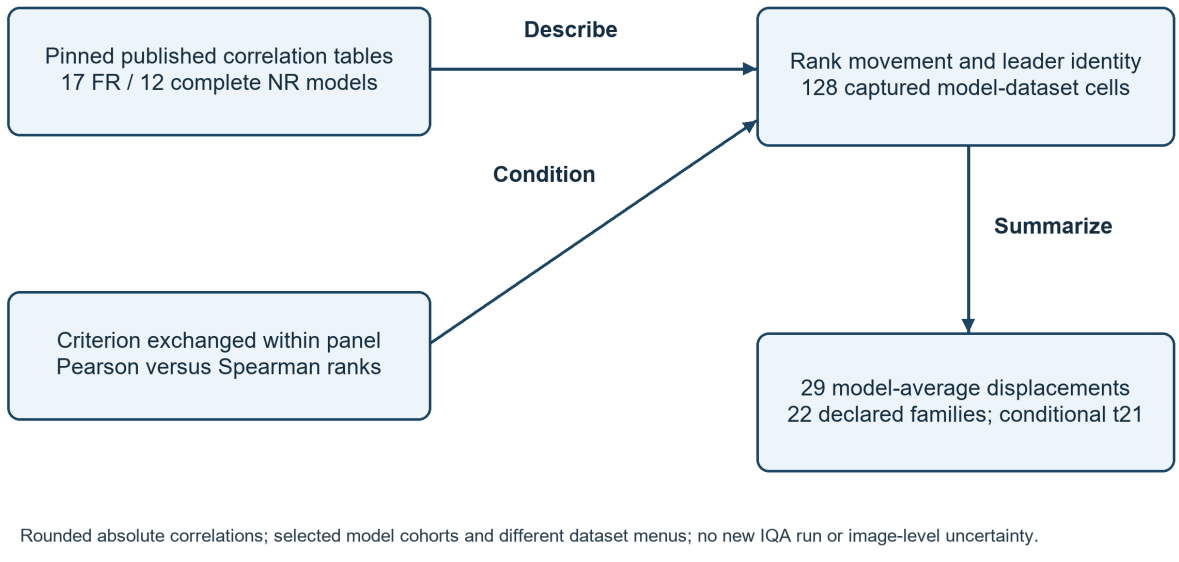

Figure 1. Fixed published cohorts and two within-panel association criteria lead to rank movement, leader checks, and conditional model-average summaries.

We cluster covariance across twenty-two declared score families, pooling SSIM spellings, LPIPS variants including STLPIPS, and the two TOPIQ entries. Other retained identifiers form separate groups.

Small-sample cluster correction and a t reference with twenty-one degrees of freedom are used, following the need to state dependence assumptions explicitly Cameron and Miller, 2015. These are conditional heterogeneity intervals under a working independent-family assumption. Common images, shared training data, and less obvious methodological ancestry can violate that assumption. The intervals do not quantify uncertainty in the source correlations.

Comparisons across datasets require careful experimental units Demšar, 2006. Here the numerical panel summaries are exact for the captured rounded records, while the regressions offer a compact description of their variation. They are not substitutes for fresh image-level resampling or controlled repeated training. No multiplicity-adjusted claim is made from the many possible model comparisons; the distinction between a descriptive scan and a family of hypothesis tests is consequential Benjamini and Hochberg, 1995. The complete analysis used local array operations Harris et al., 2020 and retained the parsed cells for inspection.

5. Widespread ordering changes, limited winner changes

The equal-model average normalized displacement is 0.114322. Its conditional family-clustered interval is 0.085909 to 0.142735. Full-reference models average 0.141544, compared with 0.075758 for the complete no-reference cohort. The latter-minus-former contrast is -0.065787, with interval -0.101680 to -0.029893. Model 2 explains 0.1821 of the observed model-average variation. These estimates describe this selected report, including its unequal dataset menus, rather than a population advantage for no-reference assessment.

Table 1. OLS summaries of normalized absolute rank displacement. *N counts model records. Covariance uses twenty-two declared family groups and t21 intervals; no image-level uncertainty is available. M1 contains an intercept; M2 adds only the cohort indicator. The fit statistic describes model-average displacement, not predictive accuracy on images.*

Model	Term	Estimate (SE)	95% CI	N
M1	Mean outcome/contrast	0.1143 (0.0137)	[0.0859, 0.1427]	29
M1: fit	R-squared = 0.0000	-	-	29
M2	FR-cohort mean displacement	0.1415 (0.0153)	[0.1098, 0.1733]	29
M2	NR minus FR table cohort	-0.0658 (0.0173)	[-0.1017, -0.0299]	29
M2: fit	R-squared = 0.1821	-	-	29

Model labels: M1 = Mean within-table criterion sensitivity; M2 = NR minus FR criterion sensitivity on different dataset menus.

Within the full-reference cohort, CSIQ has thirty-four strict pair reversals among 136 pairs and mean normalized displacement 0.191176. LIVE has twenty-six reversals and mean displacement 0.132353; TID2008 has sixteen and 0.088235; TID2013 has twenty-five and 0.154412. TOPIQ-FR remains the leader in all four panels under both criteria. Thus an unchanged winner coexists with substantial reordering below it, answering the first proposition directly.

The complete no-reference cohort contains sixty-six pairs per panel. FLIVE has three reversals, KonIQ-10k two, LIVE Challenge five, SPAQ five, and TID2013 thirteen. Corresponding normalized displacement means are 0.045455, 0.030303, 0.075758, 0.060606, and 0.166667. Only LIVE Challenge

changes leader: MUSIQ-KonIQ leads by Pearson correlation, whereas MANIQA leads by Spearman correlation. Across the nine primary panels, eight retain their leading identifier.

This pattern should not be summarized as either total stability or total fragility. A reader selecting only the first model faces fewer changes than a reader selecting several finalists or using model order as evidence of incremental progress. In particular, the TID2013 no-reference panel has more displacement than several full-reference panels despite the smaller mean for its cohort. The cohort coefficient cannot replace a panel-level account of where disagreements occur.

6. Missing reports change the recommendation problem

An available-record sensitivity calculation retains every model with a reported value in each individual panel. Full-reference results are unchanged because those records are complete. No-reference competitor counts become twenty-five for LIVE Challenge, twenty-one for KonIQ-10k, twenty-one for TID2013, twenty-three for FLIVE, and sixteen for SPAQ. These calculations answer a different question, because the model set can now vary across datasets. We report them alongside the balanced analysis rather than selecting whichever produces a more striking result.

The available-record LIVE Challenge leaders are Q-Align 8-bit under Pearson and Q-Align 4-bit under Spearman. The panel still changes leader, but the identities differ from those in the balanced twelve-model analysis. Other available-record leaders also change because additional candidates enter: for example, FLIVE selects MUSIQ-PAQ2PIQ under both criteria. These are changes in the feasible candidate set, not evidence that a retained model's underlying prediction file changed.

Data sources connected to these contexts, including the picture-quality resource Ying et al., 2019 and smartphone photography assessment Fang et al., 2020, help identify why benchmark columns should retain their dataset names and scope. They do not supply a principled missing-cell imputation. Treating every blank as a poor score would create a new ranking rule unsupported by the source. Restricting to complete records instead creates a transparent but selective population that also deserves a sensitivity display.

7. Reporting consequences and limits

A benchmark recommendation should identify its correlation criterion, candidate set, data split, score direction, and treatment of missing entries. Selecting a criterion after inspecting the desired winner risks an additional selection layer, analogous to the broader problem of evaluating a model after using the same evidence to select it Varma and Simon, 2006. The present audit does not estimate that bias. It shows the concrete reporting choices through which different recommendations can arise from the same captured source.

Rounded coefficients place a further limit on interpretation. A strict reversal is exact for the displayed numbers, but the report cannot reveal how close the unrounded quantities are or whether the difference would persist after changing images or evaluators. Absolute values also remove sign information. We therefore avoid interpreting a larger recorded magnitude as proof of correct score orientation, calibrated agreement, or statistically superior human alignment.

The twenty-two family labels are an explicit practical grouping, not a discovered independence structure. Shared datasets affect many groups simultaneously, and some nominally distinct models can share components. The exact rank and reversal summaries remain reproducible without those independence assumptions. The regression intervals are less portable and should remain secondary. Establishing general

criterion sensitivity would require a prospectively defined model population and access to paired predictions, labels, and repeated evaluation units.

The acquired record is documented with a source manifest, hashes, and exclusions in the spirit of dataset documentation Gebru et al., 2021 and reusable research information Wilkinson et al., 2016. Repository availability does not remove original dataset conditions, and the captured license text is retained rather than replaced with an assumed permissive license. This audit provides a bounded empirical reporting lesson: specify which notion of stability matters, and examine both criterion changes and coverage before turning a sorted table into a general recommendation.

8. Conclusion

Pearson and Spearman criteria produced one leader change across nine complete-cohort panels, yet altered many lower-ranked pairwise comparisons. A model-average displacement of 0.114 therefore conveys information that winner agreement alone omits. Missing reports added a separate source of recommendation change by redefining eligible competitors. These findings support a more explicit benchmark account of criterion, cohort, and decision scope; they do not establish superiority of any assessment architecture or resolve uncertainty in the underlying image-level evaluations.

References

1. 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.
2. Ding, Keyan, et al. "Image Quality Assessment: Unifying Structure and Texture Similarity." *arXiv preprint arXiv:2004.07728*, 2020. https://arxiv.org/abs/2004.07728.
3. Benjamini, Yoav and Yosef Hochberg. “Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing.” *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 57, no. 1, 1995, pp. 289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x.
4. Chen, Chaofeng; Mo, Jiadi. (2022). IQA-PyTorch: PyTorch Toolbox for Image Quality Assessment. Software repository. https://github.com/chaofengc/IQA-PyTorch/tree/d47b7ac72ed25388371d0b637076c125cdb95a3c.
5. 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.
6. 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.
7. 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.
8. 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.
9. Stevens, S. S.. (1946). On the Theory of Scales of Measurement. Science. https://doi.org/10.1126/science.103.2684.677.
10. Wang, Jianyi, et al. "Exploring CLIP for Assessing the Look and Feel of Images." *arXiv preprint arXiv:2207.12396*, 2022. https://arxiv.org/abs/2207.12396.
11. Li, Chunyi, et al. "AGIQA-3K: An Open Database for AI-Generated Image Quality Assessment" *arXiv preprint arXiv:2306.04717*, 2023. https://arxiv.org/abs/2306.04717.
12. Demšar, Janez. (2006). Statistical Comparisons of Classifiers over Multiple Data Sets. Journal of Machine Learning Research. https://www.jmlr.org/papers/v7/demsar06a.html.
13. Spearman, C.. (1904). The Proof and Measurement of Association between Two Things. The American Journal of Psychology. https://doi.org/10.2307/1412159.
14. Lin, Hanhe, et al. "KADID-10k: A Large-scale Artificially Distorted IQA Database" *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, 2019. https://doi.org/10.1109/qomex.2019.8743252.

15. Harris, Charles R.; Millman, K. Jarrod; van der Walt, Stéfan J.; Gommers, Ralf; Virtanen, Pauli; Cournapeau, David; Wieser, Eric; Taylor, Julian; Berg, Sebastian; Smith, Nathaniel J.; Kern, Robert; Picus, Matti; Hoyer, Stephan; van Kerkwijk, Marten H.; Brett, Matthew; Haldane, Allan; del Río, Jaime Fernández; Wiebe, Mark; Peterson, Pearu; Gérard-Marchant, Pierre; Sheppard, Kevin; Reddy, Tyler; Weckesser, Warren; Abbasi, Hameer; Gohlke, Christoph; Oliphant, Travis E.. (2020). Array programming with NumPy. Nature. https://doi.org/10.1038/s41586-020-2649-2.

16. Mittal, A., et al. "Making a “Completely Blind” Image Quality Analyzer" *IEEE Signal Processing Letters*, 2013. https://doi.org/10.1109/lsp.2012.2227726.

17. Bland, J. Martin, and Douglas G. Altman. "Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement." *The Lancet*, vol. 327, no. 8476, 1986, pp. 307-310. https://doi.org/10.1016/S0140-6736(86)90837-8.

18. Talebi, Hossein and Peyman Milanfar. "NIMA: Neural Image Assessment" *arXiv preprint arXiv:1709.05424*, 2017. https://arxiv.org/abs/1709.05424.

19. Varma, Sudhir and Richard Simon. “Bias in error estimation when using cross-validation for model selection.” *BMC Bioinformatics*, vol. 7, no. 1, 2006. https://doi.org/10.1186/1471-2105-7-91.

20. 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.

21. Li, Chunyi; Kou, Tengchuan; Gao, Yixuan; Cao, Yuqin; Sun, Wei; Zhang, Zicheng; Zhou, Yingjie; Zhang, Zhichao; Zhang, Weixia; Wu, Haoning; Liu, Xiaohong; Min, Xiongkuo; Zhai, Guangtao. (2024). AIGIQA-20K: A Large Database for AI-Generated Image Quality Assessment. arXiv:2404.03407. https://arxiv.org/abs/2404.03407.

22. 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.

23. 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.

24. 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.

25. 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.

26. Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. “Measurement Error in Nonlinear Models.” 2006. https://doi.org/10.1201/9781420010138.

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

28. Ke, Junjie, et al. "MUSIQ: Multi-scale Image Quality Transformer" *arXiv preprint arXiv:2108.05997*, 2021. https://arxiv.org/abs/2108.05997.

29. Lin Zhang, ; Lei Zhang, ; Xuanqin Mou, ; Zhang, D.. (2011). FSIM: A Feature Similarity Index for Image Quality Assessment. IEEE Transactions on Image Processing. https://doi.org/10.1109/tip.2011.2109730.

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