Rating Dispersion and the Difficulty of Predicting Image Quality

Gavin Fleming (TH), Vincent Anderson (SK), Nicholas Young (MN)

Abstract

Images with higher mean quality do not necessarily have more consistent ratings in the released AGIQA-3K table. Across 2,982 rows, the mean quality score and its reported rating standard deviation have a correlation of 0.083. A regression adjusted for alignment and generation condition yields a quality coefficient of -0.012, with a 95 percent interval spanning zero. This secondary analysis examines aggregate rating dispersion, not prediction errors from an image model. It shows why apparent rating difficulty should not be inferred from a mean score alone.

1. A mean hides disagreement among judgments

Two images can receive the same average quality score with very different spreads of ratings. That distinction matters when a model is assessed against aggregate human judgments. A large prediction residual might reflect a poor predictor, an ambiguous target, or several other features of the assessment process. Dispersion does not settle among these explanations, but it supplies additional descriptive information. The present study asks how the released rating spread relates to mean quality and alignment within a public score table.

Decoupling Semantics from Distortions proposes a multiscale two-stream visual-language assessment approach (Meng, 2026). Its separation of evidence motivates attention to how a human target is defined and how stable it appears. The present analysis does not run that model or assess its prediction difficulty. It analyzes published mean scores and standard deviations. The distinction is essential: a relationship involving rating spread cannot be narrated as an observed relationship involving model residuals that were never computed.

The choice of statistical unit should precede the choice of a standard error. A dataset can contain many rows but relatively few independently acquired histories. Heteroskedasticity-consistent covariance protects an otherwise specified regression against unequal residual variances (White, 1980). It does not automatically accommodate observations sharing an environment, subject, or recording. Cluster-robust inference makes that dependence part of the variance calculation (Cameron and Miller, 2015). The grouping must follow the observation process and the claim being made; an identifier that is convenient to store is not necessarily the relevant independent unit. For a descriptive analysis, preserving the groups also allows readers to recalculate a different population average. An uncertainty interval remains conditional on the observations and assumptions actually available.

The observational units in a generated-image score table include prompts, conditions, and individual images. AGIQA-3K makes those distinctions relevant to secondary analysis (Li et al., 2023). Several rows can share a prompt without constituting repeated draws from every underlying generation process. A careful study should identify what was actually released and avoid inferring an unobserved sampling design. Datasheets motivate preserving the purpose, provenance, and limitations of the collection (Gebru et al., 2021). Documented software procedures make variable construction and assessment splits easier to reproduce (Pedregosa et al., 2011). These practices are especially useful when only scores and metadata are processed. Such a study can establish associations and coverage properties of the published table while leaving image-level mechanisms untested. Its results should name the outcome scale and weighting population, retain incomplete combinations, and distinguish descriptive labels from known experimental

controls. The absence of image processing is not a defect if the question concerns the score record; it simply defines which conclusions the evidence can support.

2. Released aggregates and the fitted specification

The input contains 2,982 image rows associated with 300 prompts and ten generation-condition labels. The analysis uses the released quality mean, alignment mean, quality-rating standard deviation, and alignment-rating standard deviation. Condition labels are derived consistently from source filenames. All calculations operate on the public score record. Individual participant responses are unavailable in this analysis, and no new ratings are collected. Consequently, the study cannot reconstruct rater-specific behavior or estimate each mean's sampling variance from the standard deviation alone.

The first calculation describes the average reported spread and the correlation between quality mean and quality spread. A conventional ordinary least squares model then relates quality-rating standard deviation to quality mean, alignment mean, and condition indicators. Inference uses prompt-clustered standard errors with the saved t-based intervals. The coefficients are partial associations in the observed table. They do not identify why raters disagree or establish a causal effect of changing perceived quality on response dispersion.

Mean reported quality spread is 0.600, while mean alignment spread is 0.741. These quantities refer to different judgment dimensions and should not automatically be interpreted as a contest in reliability. The score scales, task instructions, and rater process matter to that comparison. The quality mean has a Pearson correlation of 0.083 with quality spread. This weak pooled relationship already suggests that high or low mean quality does not provide a simple ranking of rating disagreement across the released rows.

A proposed explanation should be distinguishable from alternative explanations using observations the study can actually collect. Diagramming the relationships among selection, exposure, measurement, and outcome makes that requirement explicit (Greenland et al., 1999). The diagram does not establish its own correctness, but it prevents adjustment choices from remaining invisible. The same transparency is needed when investigators inspect many endpoints before choosing a narrative. Controlling false discovery concerns the behavior of a specified testing family (Benjamini and Hochberg, 1995); it is not a general certificate of scientific validity. Primary endpoints, secondary descriptions, and future mechanism tests should be named separately. Readers can then assess whether an attractive pattern was predicted, explored, or merely noticed after the available comparisons were examined.

Visual-language alignment and image quality have related but distinguishable measurement foundations. CLIP connects images and text through language-supervised representation learning (Radford et al., 2021). Its similarities can inform an assessment, but they do not automatically define perceptual quality or an application's complete notion of semantic correctness. MUSIQ develops a multiscale image-quality approach (Ke et al., 2021), NIMA studies learned aesthetic assessment (Talebi and Milanfar, 2017), and NIQE offers a natural-scene-statistics-based quality measure (Mittal et al., 2013). These methods illustrate different targets and information assumptions rather than interchangeable instruments. A generated image can look convincing while missing the instruction, or satisfy a prompt while containing visible artifacts. The evaluation should retain those possibilities instead of treating every score as a noisy observation of one universal quality. In a secondary analysis of released human judgments, relationships between quality and alignment describe those judgments. They do not establish how one of these automated instruments would perform unless its predictions are actually obtained and assessed under a suitable independent protocol.

3. Adjustment does not turn dispersion into a mechanism

In the adjusted regression, the quality-mean coefficient is -0.012, with a 95 percent interval from -0.030 to 0.005. The alignment-mean coefficient is 0.009, with an interval from -0.007 to 0.025. Both intervals include zero under this specification. The reported coefficient of determination is 0.029. The model therefore leaves most observed variation in the aggregate quality spread undescribed by these particular covariates; that is a statement about this fit, not proof that the remaining variation is inherently unpredictable.

Rating spread is also different from uncertainty about the population mean. Converting a standard deviation into a standard error would require the relevant rating count and assumptions about dependence and sampling. The present table analysis does not invent those inputs. Nor does a large spread necessarily indicate annotation failure: observers may apply different legitimate preferences to an image. Establishing the cause would need participant-level evidence, a revised task, or a specifically designed follow-up assessment rather than interpretation of aggregate dispersion alone.

For future quality-model evaluation, the practical implication is to retain both the target mean and the documented information about its spread. A study could test whether model errors vary with dispersion after obtaining independent predictions and defining its assessment split. That experiment has not been performed here. The present result supports a more limited conclusion: the released mean score is an incomplete description of the judgment target, and a simple monotonic relationship between quality level and rating agreement should not be assumed from these data.

Established quality datasets and metrics remind the analyst that observation populations differ. KonIQ-10k studies quality in a large in-the-wild image collection (Hosu et al., 2019), while KADID-10k supports assessment of declared image distortions (Lin et al., 2019). These are not automatically the same population as generated imagery. Structural similarity (Wang et al., 2004) and learned perceptual similarity (Zhang et al., 2018) also require an appropriate interpretation of the compared images. FID instead compares distributions in a learned representation (Heusel et al., 2017). None of these objects is interchangeable with an individual human alignment judgment. A study should state which references or populations a metric needs and whether those inputs exist in the experiment. Distributional quality cannot be inferred solely from one image's mean score, and a high pairwise resemblance can reward repetition rather than successful instruction execution. The relevant lesson for a score-only analysis is to keep its outcome boundary precise. It can describe published judgments and propose a future metric comparison without reporting that the comparison has already been performed.

A validation split defines which form of transfer has been observed. Related observations can share subjects, scenes, prompts, or acquisition conditions, so randomly splitting rows may leave important information on both sides of the boundary. Structured cross-validation addresses these dependencies (Roberts et al., 2017). Model selection introduces another boundary: choosing a representation, threshold, or tuning parameter using the final assessment data makes the resulting error estimate optimistic (Varma and Simon, 2006). The appropriate held-out unit follows the intended future use. Generalization to a new frame, a new trajectory, and a new subject are not equivalent claims. A study should record those units before fitting or choosing a scoring rule, and preserve an assessment set that has not become part of iterative development.

Difficult cases can disappear at several stages: data eligibility, generation, measurement, and final scoring. Each exclusion changes the population represented by the reported result. Coverage-aware evaluation makes those decisions visible (Geifman and El-Yaniv, 2017). A confidence score should also be checked when the input distribution changes, because uncertainty under shift is an empirical property

rather than a guarantee of the model interface (Ovadia et al., 2019). An informative record preserves failures to produce an output and failures to obtain a valid measurement. Treating either as absent data without explanation can inflate apparent reliability. A proposed method should state its unsupported conditions and test the boundary between accepted and rejected cases as carefully as the quality of its successful outputs.

4. Measurement, transfer, and reporting

Preference-based evidence can complement quality and alignment scores without replacing their definitions. Pick-a-Pic records comparative choices (Kirstain et al., 2023), while Human Preference Score studies a preference-oriented measurement route (Wu et al., 2023). DINO's self-supervised features (Caron et al., 2021) offer another representation of visual content. Subject-conditioning approaches including DreamBooth (Ruiz et al., 2022) and IP-Adapter (Ye et al., 2023) show that generation can also differ in the external evidence it receives. A fair assessment should retain those differences in the experiment record. One workflow's advantage can arise partly from richer information, and one score's usefulness can depend on access to judgments that another task lacks. The analysis should therefore state what is observed at prediction or assessment time, not merely the names of the models discussed. In a released-score study, no new preference or conditioning performance follows from those citations. The direct evidence remains the published human outcomes and the explicitly computed relationships among them, with additional validation treated as a separate empirical undertaking.

A substitute measurement needs evidence about its relationship to the intended judgment. Two scores may correlate strongly while differing systematically in particular conditions. Method-comparison analysis distinguishes association from agreement (Bland and Altman, 1986). When a system additionally reports confidence, calibration asks whether those probabilities correspond to observed outcome frequencies (Guo et al., 2017). These are separate properties: a metric can rank examples usefully while having a distorted numerical scale or unreliable confidence. An evaluation should therefore retain the original target observations when possible, describe the scale of each endpoint, and examine disagreements instead of averaging them away. A convenient automated score becomes more credible when its limitations are measured in the conditions that motivate its use.

An evaluation remains comparable only if its inputs and transformations can be identified. The FAIR principles emphasize persistent identification and reusable scientific metadata (Wilkinson et al., 2016). Work on machine-learning technical debt highlights how dependencies and hidden couplings accumulate in a pipeline (Sculley et al., 2015). In this setting, a score should be linked to the exact references, generation settings, evaluator version, and aggregation rule that produced it. A replacement checkpoint or a revised eligibility filter may change the scientific question even when the output column keeps the same name. Preserving such changes in a release record permits later comparisons to be interpreted honestly. The objective is not maximum documentation volume, but a recoverable account of the decisions that affect the result.

5. Conclusion

Quality-rating dispersion has little pooled association with mean quality in the released table, and the adjusted coefficients remain uncertain under the chosen specification. This score-based finding does not measure model error or explain rater disagreement. It motivates keeping target spread visible in future evaluation and obtaining the additional observations needed before uncertainty, annotation difficulty, and automated prediction difficulty are treated as equivalent properties of the same generated image.

References

Meng, Zijie. "Decoupling Semantics from Distortions: Multi-Scale Two-Stream Vision-Language Alignment for AI-Generated Image Quality Assessment." *arXiv preprint arXiv:2606.16799* (2026).

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.

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.

Greenland, Sander, Judea Pearl, and James M. Robins. “Causal Diagrams for Epidemiologic Research.” *Epidemiology*, vol. 10, no. 1, 1999, pp. 37-48. https://doi.org/10.1097/00001648-199901000-00008.

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.

Roberts, David R., et al. “Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure.” *Ecography*, vol. 40, no. 8, 2017, pp. 913-929. https://doi.org/10.1111/ecog.02881.

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.

Wilkinson, Mark D., et al. "The FAIR Guiding Principles for Scientific Data Management and Stewardship." *Scientific Data*, vol. 3, 2016, article 160018. https://doi.org/10.1038/sdata.2016.18.

Sculley, D., et al. "Hidden Technical Debt in Machine Learning Systems." *Advances in Neural Information Processing Systems*, vol. 28, 2015. https://papers.nips.cc/paper/2015/hash/86df7dcfd896fcaf2674f757a2463eba-Abstract.html

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.

Guo, Chuan, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. “On Calibration of Modern Neural Networks.” *Proceedings of the 34th International Conference on Machine Learning*, PMLR, vol. 70, 2017, pp. 1321-1330. https://proceedings.mlr.press/v70/guo17a.html.

Geifman, Yonatan, and Ran El-Yaniv. “Selective Classification for Deep Neural Networks.” *Advances in Neural Information Processing Systems*, vol. 30, 2017. https://arxiv.org/abs/1705.08500.

Ovadia, Yaniv, et al. “Can You Trust Your Model’s Uncertainty? Evaluating Predictive Uncertainty Under Dataset Shift.” *Advances in Neural Information Processing Systems*, vol. 32, 2019. https://arxiv.org/abs/1906.02530.

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.

Radford, Alec, et al.. "Learning Transferable Visual Models From Natural Language Supervision" *arXiv preprint arXiv:2103.00020*, 2021. https://arxiv.org/abs/2103.00020.

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

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.

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.

Wang, Zhou, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. "Image Quality Assessment: From Error Visibility to Structural Similarity." *IEEE Transactions on Image Processing*, vol. 13, no. 4, 2004, pp. 600-612. https://doi.org/10.1109/TIP.2003.819861.

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.

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

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

Kirstain, Yuval, et al.. "Pick-a-Pic: An Open Dataset of User Preferences for Text-to-Image Generation" *arXiv preprint arXiv:2305.01569*, 2023. https://arxiv.org/abs/2305.01569.

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.

Caron, Mathilde, et al.. "Emerging Properties in Self-Supervised Vision Transformers" *arXiv preprint arXiv:2104.14294*, 2021. https://arxiv.org/abs/2104.14294.

Ruiz, Nataniel, et al.. "DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation" *arXiv preprint arXiv:2208.12242*, 2022. https://arxiv.org/abs/2208.12242.

Ye, Hu, et al.. "IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models" *arXiv preprint arXiv:2308.06721*, 2023. https://arxiv.org/abs/2308.06721.

Heusel, Martin, et al.. "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium" *arXiv preprint arXiv:1706.08500*, 2017. https://arxiv.org/abs/1706.08500.

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

Pedregosa, Fabian, et al. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research*, vol. 12, 2011, pp. 2825-2830. https://jmlr.org/papers/v12/pedregosa11a.html.