Prompt Style and Adjusted Quality Judgments in AGIQA-3K

Oscar Pearson (FM), Christian Nelson (PL), Joseph Newman (IR)

Abstract

Baroque and realistic style labels have higher adjusted quality means than the abstract-style reference category in a secondary analysis of AGIQA-3K. The estimated differences are 0.194 and 0.209 score units after accounting for alignment, generation condition, and recorded adjective count. These are associations in 2,982 released score rows, with uncertainty clustered over 300 prompts. They do not show that changing style causes a quality gain. The findings motivate separating semantic and perceptual evidence when interpreting generated-image quality assessments.

1. A style label can describe several correlated properties

Style instructions can change the intended appearance of an image and the expectations under which its quality is judged. A difference in mean quality across labels may therefore combine content, generation conditions, and evaluation conventions. This study examines the released AGIQA-3K score table rather than image pixels. Its question is whether recorded style categories remain associated with quality after a small set of available score and condition variables are included in a conventional linear regression.

The two-stream assessment proposed in Decoupling Semantics from Distortions motivates keeping semantic correspondence and visual quality conceptually distinct (Meng, 2026). Its multiscale visual-language architecture is not reproduced here. Instead, human alignment means enter a score-based descriptive model. This is a narrower empirical contribution: it asks how quality judgments vary within a documented table, without claiming that the fitted style coefficients reveal the behavior or comparative accuracy of the focal image-quality model.

Before fitting a score model, an analyst should reconstruct the table's observation process. AGIQA-3K provides human-assessment information for generated images (Li et al., 2023), but a released mean score is not a new independent measurement each time it is reused in another specification. Provenance and intended-use documentation help define the scientific role of those rows (Gebru et al., 2021). Reproducible software choices matter for grouping, preprocessing, and validation (Pedregosa et al., 2011). A score-only analysis should therefore state that it processes the published judgments rather than the underlying visual files. It should preserve prompt and condition identifiers, distinguish absent combinations from missing values within present rows, and describe how any derived categories were obtained. These records permit another reader to reconstruct the comparison and its denominator. They also prevent familiar benchmark terminology from implying stronger access than the study has. An association among aggregate human scores remains useful evidence about the released table even when participant-level uncertainty, image features, and new generation outcomes are unavailable for a broader interpretation.

The representation used for scoring can influence which differences are visible. CLIP's language-supervised features (Radford et al., 2021) emphasize a particular connection between text and imagery. Multiscale quality modeling in MUSIQ (Ke et al., 2021), aesthetic learning in NIMA (Talebi and Milanfar, 2017), and natural-scene-statistics modeling in NIQE (Mittal et al., 2013) address other aspects of visual assessment. Each provides an example of why the term quality needs a scientific definition. A favorable semantic score does not establish freedom from perceptual distortion, and a favorable perceptual score does not establish that a detailed instruction was followed. For generated imagery, both dimensions can be relevant at once. The analysis should state whether its outcome concerns human judgments, an automated prediction, or a task-specific decision. Correlations across those objects can

support a useful measurement study, but their meaning depends on the observation population and evaluation procedure. Without obtaining a model's outputs, a released-score analysis should remain about the available judgments rather than implying a reproduction or validation of the model itself.

2. Data construction and adjusted contrasts

The analysis uses all 2,982 released image rows and their 300 prompt identifiers. Generation condition is derived from the source filename by removing its final image index. There are ten such condition labels. Empty style fields are represented by an unspecified category. The adjective-count covariate counts nonmissing entries in the two released adjective fields; it is not a natural-language estimate of all descriptive content in a prompt. No new images, ratings, or text annotations are collected.

Ordinary least squares relates mean quality to style indicators, condition indicators, mean alignment, and the recorded adjective count. Abstract style is the reference category. Standard errors are clustered by prompt, with the saved confidence intervals using the fitted cluster procedure and a t reference. This accounts for a declared source of within-prompt dependence. It does not remove unmeasured content differences or make the style labels randomly assigned. The model also does not include a separate fixed effect for every prompt.

Unadjusted mean quality is 2.332 for abstract style, 2.442 for anime, 2.631 for baroque, 2.569 for realistic, 2.490 for sci-fi, and 2.507 for unspecified style. The category counts range from 277 to 280 for the specified labels, while unspecified style has 1,587 rows. Those unequal counts matter to the table's overall composition. A pooled score average therefore represents the observed mixture rather than equal representation of the six recorded categories.

Replication has several levels: a second measurement, a second trajectory, and a second independently acquired setting are different additions to evidence. Statistical calculations should retain those distinctions. Heteroskedasticity-robust regression inference concerns unequal conditional variances (White, 1980); cluster-robust inference concerns dependencies shared within groups (Cameron and Miller, 2015). The group definition should be justified through the data collection and evaluation process, and not chosen because it produces a preferred significance result. Unequal group sizes also influence what a pooled average represents. Reporting both the group count and observation count permits a reader to see whether the calculation describes common cases, heavily recorded cases, or an explicitly balanced group population. No variance estimator alone resolves that choice of estimand.

The scientific interpretation should specify what would change under the proposed intervention. Holding an observed covariate fixed statistically is not equivalent to manipulating it in an environment. A causal graph helps separate confounders, mediators, and selection mechanisms (Greenland et al., 1999). It can also reveal why controlling for every recorded variable is not automatically defensible. Meanwhile, an evaluation with multiple conditions and endpoints creates a family of opportunities for apparently striking findings. Multiplicity methods require a declared set of comparisons and an appropriate inferential purpose (Benjamini and Hochberg, 1995). Descriptive tables may be useful without testing every cell. When the data do not identify a mechanism, the appropriate contribution can be a sharper question and a transparent comparison rather than an unsupported causal answer.

Transfer across image populations requires evidence rather than a familiar metric name. KonIQ-10k's in-the-wild setting (Hosu et al., 2019) differs from the controlled distortion emphasis of KADID-10k (Lin et al., 2019). Structural similarity (Wang et al., 2004) and LPIPS (Zhang et al., 2018) assess particular forms of pairwise correspondence, while FID (Heusel et al., 2017) summarizes a feature-distribution comparison. Their differing inputs and units make them complementary tools, not alternative spellings of

the same outcome. For generated images, reference availability and semantic intent add further constraints. A reference-based resemblance can be high even when a requested change is absent, and a distributional score need not identify which individual images fail an instruction. The assessment should therefore connect each instrument to a declared claim. Secondary analysis of human score metadata can inform that design, but it does not supply missing images, features, or predictions. The bounded empirical result remains the association observed in the released judgments, with any proposed metric evaluation clearly identified as future work.

3. What remains after the available adjustment

The adjusted baroque contrast is 0.194, with a 95 percent interval from 0.068 to 0.320. The realistic contrast is 0.209, with an interval from 0.088 to 0.330. Anime has an estimated contrast of 0.061 and an interval spanning zero, from -0.060 to 0.181. Sci-fi and unspecified style have positive estimated contrasts of 0.134 and 0.139. The regression's reported coefficient of determination is 0.760, describing fit to the observed rows rather than held-out predictive performance.

These coefficients should not become a recommendation to change every prompt to a particular style. Alignment is itself an assessed outcome and may share determinants with quality. Conditioning on it defines a partial association rather than isolating an intervention. The table also provides limited information about generation settings beyond its labels. The comparison cannot establish whether a difference arises from a style instruction, its associated content, a generation configuration, or a judgment process that varies with the intended appearance.

A future image-level study could examine matched prompts and independently validate the distinction between semantic and distortion evidence. That would require actual model predictions and appropriate assessment splits. The present results instead identify heterogeneity that a quality model's evaluation should consider. Reporting style-specific performance and uncertain cases could reveal whether an aggregate advantage is concentrated in familiar categories. The source table supports that motivation, but it does not supply an unperformed comparison between competing quality architectures or new evidence about their internal representations.

The protected assessment population should be defined before investigators inspect detailed failures. Once a collection of cases guides repeated revisions, it has become development information even if its filenames still carry an evaluation label. Cross-validation for dependent data treats the grouping of observations as a substantive design decision (Roberts et al., 2017). Research on model selection likewise shows why tuning and final error estimation should not reuse the same information (Varma and Simon, 2006). A practical record therefore distinguishes fitting data, diagnostic examples, validation data, and the final external assessment. Where only a fixed public table is available, a descriptive analysis remains legitimate, but it should not be renamed a held-out prediction experiment without generating and assessing predictions under a protected split.

A score is useful only relative to the question it is supposed to answer. Correlation measures a relationship between varying quantities, whereas agreement concerns the discrepancies produced when one procedure substitutes for another (Bland and Altman, 1986). Confidence estimates add a further question about calibration (Guo et al., 2017). These distinctions matter when a learned evaluator becomes the principal source of evidence about another learned system. The evaluation should preserve examples with conflicting scores and obtain independent target judgments where feasible. Numerical ranges and normalization choices should also be documented. An apparently small difference may have a different practical meaning near a decision threshold than in an unimportant part of the scale. Reporting the intended consequence gives those comparisons an interpretable purpose.

4. Measurement, transfer, and reporting

An image that receives a high quality judgment need not win every user comparison. Pick-a-Pic (Kirstain et al., 2023) and Human Preference Score (Wu et al., 2023) motivate attention to preference as a distinct assessed outcome. DINO (Caron et al., 2021) provides self-supervised visual features that can characterize content under another representation. DreamBooth (Ruiz et al., 2022) and IP-Adapter (Ye et al., 2023) illustrate how subject or image references can influence the generation task before scoring begins. These sources help explain why an evaluation should specify both its input evidence and its target judgment. A method with additional references or adaptation opportunities may solve a different information problem. Likewise, a predictor of aggregate alignment may require human information unavailable in a practical no-reference quality system. The report should make such access explicit. A secondary score-table analysis remains informative when it describes its available outcomes precisely and reserves statements about preference prediction, subject conditioning, or visual-feature performance for studies in which those outputs were actually obtained and independently assessed.

A useful reliability claim contains both a performance statement and a scope statement. Selective prediction shows why a system that answers only some cases must report how many it accepts (Geifman and El-Yaniv, 2017). Confidence estimation under a changed data distribution requires independent checking (Ovadia et al., 2019). These requirements apply to generators and to the automated judges used to assess them. If a judge cannot evaluate a profile, an occluded object, or an unusual interaction, that inability should remain visible rather than removing the case from the generator's score silently. Recording the entire decision path permits an analyst to identify whether an apparent improvement came from better outputs, narrower eligibility, or a more restrictive measurement procedure.

Reproducibility begins with preserving which scientific objects are being linked. Stable identifiers and reusable metadata help connect inputs, reference observations, model versions, and reported outcomes (Wilkinson et al., 2016). The scoring pipeline also has dependencies that can change the interpretation of a result: preprocessing, representation checkpoints, thresholds, and aggregation code can all matter (Sculley et al., 2015). A compact release should retain those decisions rather than only the final mean. Where source data cannot be redistributed, an access description and hashes can still identify the analyzed inputs. This record allows later users to distinguish a change in the underlying system from a change in how the system was measured. Reproducibility is therefore part of interpretation, not simply a final packaging step.

5. Conclusion

Recorded style categories are associated with human quality means after the available adjustment, with positive baroque and realistic contrasts relative to abstract style. The findings are descriptive and depend on the released score population. Their practical contribution is to motivate style-aware reporting alongside separate semantic and perceptual endpoints. They establish neither a causal style effect nor a reproduction of the focal two-stream assessment model, both of which would require additional evidence beyond the score table.

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.