When Image Quality and Prompt Alignment Diverge: A Secondary Analysis of AGIQA-3K

Anthony Barnes (WS), Jeffrey Green (TN), Calvin Robertson (SS)

Abstract

AGIQA-3K gives two judgments for the same generated images: perceptual quality and alignment with the prompt. Across 2,982 released records their Pearson correlation is 0.814, but the quality-on-alignment slope falls from 0.813 to 0.530 after accounting for generation condition and prompt attributes. Twenty-four images also occupy opposite outer quartiles of the two rankings. The outcomes are strongly related without being interchangeable. A regression analysis of these public scores clarifies what evidence motivates separate quality and semantic assessments, while leaving the performance of any particular two-stream model to direct testing.

1. The same image can answer two different questions

An image can look polished while omitting a requested object. It can also satisfy the main nouns in a prompt while containing awkward texture, inconsistent lighting, or implausible local structure. A quality assessment therefore needs to distinguish the visual experience from the image's relationship to a requested description. Treating those questions as separate does not imply that their answers must be statistically independent. Better generators, easier prompts, or shared human preferences can move both outcomes in the same direction.

Decoupling Semantics from Distortions develops MST-CLIPIQA, a multi-scale two-stream approach to this assessment problem (Meng, 2026). Its reported design separates coarse semantic information from finer visual information, then combines streams through gated fusion. Prompt correspondence can be incorporated when a prompt is available. The present analysis examines an empirical premise relevant to that design: whether released human quality and alignment scores convey distinct information after basic features of their generation context are considered. It does not execute the architecture, reproduce its benchmark gains, or establish that two encoders are the best implementation of the distinction.

The data come from AGIQA-3K, an openly released database developed for AI-generated image quality assessment (Li et al., 2023). The official repository distributes a compact table of prompts, generation labels, and aggregated subjective scores. Those scores allow an actual secondary analysis without downloading images or training a predictor. The resulting evidence concerns the joint distribution of published judgments, not a new assessment by human participants.

2. Records, conditions, and the analysis boundary

The downloaded table contains 2,982 unique image filenames and 300 distinct prompt strings. Ten generation conditions are recovered from the documented filenames by removing the terminal image index. They include normal and modified settings within some generator families, so the term condition is more precise than calling all ten independent models. The supplied style and adjective fields describe aspects of the prompt. Unspecified style is retained as a category rather than used as a reason to delete an image.

Both outcome columns are complete. Perceptual quality averages 2.5004 and alignment averages 2.5000 on their released normalized scales. Similar overall means are not evidence that they measure the same construct; normalization and aggregation can produce similar locations. The accompanying standard-deviation columns summarize source judgments, but the individual rating records are not

present. Consequently, this study does not reconstruct rater effects, test-retest reliability, or the sampling error of each image's mean score.

The nominal ten-by-300 condition-prompt grid would contain 3,000 records. The archive contains eighteen fewer. Counts range from 290 to 300 by condition. The analysis preserves all observed rows and performs no synthetic completion of the grid. It does not infer why particular entries are absent. Dataset documentation is essential for distinguishing a released observation from a combination that an analyst merely expects to exist (Gebru et al., 2021). The original files, their hashes, and the transformations into analysis variables remain available for checking.

This boundary also matters for reproducibility. Scikit-learn illustrates the importance of stable analysis interfaces and documented software in machine-learning research (Pedregosa et al., 2011); here the statistical calculations use ordinary regression and correlation routines with saved inputs. The purpose is a transparent small-data reanalysis. No image embeddings, text-image similarity scores, pretrained checkpoints, or newly generated examples enter the calculation.

3. Three regressions that answer progressively narrower questions

The first model regresses perceptual quality on alignment alone. Its coefficient describes the pooled association across the released images. The second adds fixed effects for the ten generation conditions, categorical prompt style, and the number of nonmissing adjective fields. It compares images after removing those observed differences in average scores. The third adds prompt fixed effects and condition fixed effects. Because prompt identity already absorbs its fixed attributes, separate style and adjective controls are unnecessary in that specification.

All three use ordinary least squares. Their purpose is descriptive decomposition, not deployment prediction or causal estimation. Images generated from the same prompt may share difficulty and content, so standard errors are clustered by the 300 prompt strings, with a finite-cluster adjustment and a t reference distribution (Cameron and Miller, 2015). A heteroskedasticity-consistent covariance estimator addresses a different problem and would not by itself capture this prompt grouping (White, 1980). The numerical intervals are conditional on the released aggregate scores; they do not incorporate the unavailable uncertainty of the underlying individual ratings.

Condition adjustment can change a slope because generator differences contribute to the pooled covariance. It does not identify a manipulation that changes alignment while holding every visual property fixed. A causal graph would also need image content, model settings, selection, and rating processes before supporting such an intervention interpretation (Greenland et al., 1999). The reported regressions therefore describe co-movement. They do not estimate how much improving semantic alignment would improve perceptual quality.

Two complementary summaries avoid relying entirely on a linear relationship. Pearson and Spearman correlations summarize linear co-movement and rank association. Empirical percentile ranks identify images simultaneously in the top quartile of one outcome and bottom quartile of the other. These counts are descriptive flags. They are not newly discovered annotation errors, and the images themselves were not visually inspected for this analysis.

4. Strong co-movement leaves room for different judgments

The Pearson correlation is 0.8141 and the Spearman correlation is 0.7419. Thus the two outcomes share substantial variation, and their rankings are related. A claim that semantics and quality are unrelated would contradict these data. The relevant question is whether one outcome can serve as an interchangeable substitute for the other, which requires much more than positive correlation.

Quality-on-alignment specification	Slope	95 percent interval	In-sample R-squared
Unadjusted	0.813	0.791 to 0.836	0.663
Condition, style, adjectives	0.530	0.478 to 0.582	0.760
Condition and prompt fixed effects	0.585	0.528 to 0.642	0.810

The adjusted slope is smaller than the pooled slope but remains positive. The prompt-fixed-effect result also remains positive. Accordingly, part of the pooled relationship reflects differences associated with observed generation context, while the available controls do not remove the relationship entirely. These coefficients are in released score units. They should not be read as standardized effect sizes or as percentage improvements from an algorithm.

Differences among condition means illustrate why this adjustment matters. The normal Midjourney condition has mean quality 3.626 and mean alignment 3.310, whereas its low-step condition has means 2.387 and 2.815. This is a comparison between the observed released groups, not a newly randomized intervention on sampling steps. It nevertheless shows that the two judgment dimensions can move by different amounts across a named condition contrast. Complete condition summaries are saved rather than restricted to this illustrative pair.

There are nineteen images in the top quality quartile and bottom alignment quartile, and five in the top alignment quartile and bottom quality quartile. These twenty-four records are a small fraction of the database, so the result should not be described as widespread opposition between the outcomes. They demonstrate a narrower point: high rank on one does not guarantee high rank on the other, even within a dataset showing strong overall correlation.

As a sensitivity calculation, each condition is omitted in turn and the second regression is refitted. The alignment coefficient ranges from 0.502 to 0.561. The positive adjusted relationship is therefore not dependent on one retained generation condition. This is a leave-condition-out descriptive exercise, not a held-out prediction experiment. No model is selected by this range, and the omitted-condition fits are not reported as independent replications.

5. Why a separate semantic pathway is plausible but unproven here

CLIP demonstrates the usefulness of learning visual representations through language supervision (Radford et al., 2021). A representation useful for matching text and images need not retain every local degradation important to a quality judgment. Conversely, a texture-sensitive representation can respond to local artifacts without understanding which requested entities and relationships are missing. MST-CLIPIQA's separation is a proposed architectural response to that tension. The current score analysis supports preserving the distinction in the target labels; it cannot determine whether the proposed representation successfully achieves it.

Other quality research highlights complementary aspects of measurement. MUSIQ explicitly addresses multiple image scales (Ke et al., 2021), while NIMA predicts distributions of image assessments (Talebi and Milanfar, 2017). NIQE evaluates natural-image regularities without a reference image (Mittal et al., 2013). These are different choices about information, supervision, and the endpoint being approximated. A two-stream assessment should be compared with appropriate single-stream and multi-scale baselines under matched data access, rather than justified only by the existence of two label columns.

The evaluation population is equally important. KonIQ-10k emphasizes quality variation in a broad collection of real-world photographs (Hosu et al., 2019), whereas KADID-10k supplies controlled artificial distortions (Lin et al., 2019). AI-generated imagery may contain content failures and unfamiliar artifact combinations in addition to those distortion families. Transfer across those populations requires an explicit test. Performance on one does not establish that a quality scale or a learned proxy has unchanged meaning on the others.

Pairwise image metrics supply another distinction. Structural similarity uses a reference comparison (Wang et al., 2004), and learned perceptual similarity uses deep visual features (Zhang et al., 2018). They are not the same task as predicting human quality from an image alone or checking the image against a text prompt. Distributional evaluation associated with generative modeling, including the Fréchet Inception Distance (Heusel et al., 2017), asks yet another question. The public AGIQA table cannot be used to calculate these image-based metrics without the images and the relevant implementation.

6. Preferences, personalization, and the meaning of a score

Human preference is not automatically synonymous with either quality or prompt alignment. Pick-a-Pic collects user preferences between generated alternatives (Kirstain et al., 2023), and Human Preference Score v2 develops a benchmark around human preferences for synthesized images (Wu et al., 2023). A preference may integrate style, content, novelty, and other considerations. A future assessment study should specify whether it predicts a preference, a quality rating, or correspondence to a prompt before comparing correlations across benchmarks.

The conditioning information also changes the target. DreamBooth addresses subject-driven generation through model adaptation (Ruiz et al., 2022), while IP-Adapter adds an image-conditioning route compatible with text (Ye et al., 2023). Subject fidelity is meaningful only when an appropriate reference is available. A prompt-only alignment score cannot establish that a generated subject retains a particular appearance. DINO's self-supervised visual representations offer another possible measurement component (Caron et al., 2021), but using them would require testing the resulting score against the intended judgment rather than assuming generic visual features define it.

The method-comparison lesson is that correlation does not establish agreement (Bland and Altman, 1986). In this study the two columns are deliberately different judgments, so agreement should not be demanded by default. The practical danger is using one as a proxy for the other without documenting what is lost. The observed rank disagreements identify candidate cases for a subsequent visual and rating study, not a set of examples whose interpretation is already known.

7. From descriptive association to a useful predictor

The reported R-squared values are fitted to the same observations used for estimation. They provide no out-of-sample performance claim. A predictor intended for unseen prompts should separate prompt groups across training and evaluation (Roberts et al., 2017), and any choice of architecture or tuning parameters must be protected from the final test set (Varma and Simon, 2006). A leave-generator or leave-condition-out prediction experiment could answer a further transfer question, but it would need actual predicted outcomes on the omitted data.

Uncertainty should be evaluated as an output in its own right. Calibration connects forecast confidence with observed outcomes (Guo et al., 2017), while distribution shift can undermine familiar uncertainty estimates (Ovadia et al., 2019). A system that abstains from difficult images should report its coverage as well as performance on accepted cases (Geifman and El-Yaniv, 2017). If a large collection of datasets, conditions, and output dimensions is compared, multiplicity should be handled transparently (Benjamini and Hochberg, 1995). These requirements concern future model evaluation; they are not accomplishments inferred from the present regressions.

Reproducibility includes retaining which prompt fields were available to a model. Stable identifiers and reusable metadata support that record (Wilkinson et al., 2016). Versioned data, preprocessing, and scoring dependencies are also important because apparently minor pipeline changes can alter an evaluation's interpretation (Sculley et al., 2015). Here the released aggregate-score table and all transformations are saved, allowing the correlations, fixed-effect specifications, and rank counts to be recalculated directly.

8. Conclusion

AGIQA-3K's quality and alignment scores are strongly associated, but the association changes after observed generation context is considered and does not prevent opposing rank positions for some images. A useful assessment should preserve both targets rather than select one through convenience. The data motivate testing architectures that distinguish semantic and perceptual information, including the two-stream proposal discussed here. Whether such a model improves measurement remains an empirical question for matched training, protected evaluation, and direct comparison with the intended human judgments.

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.