Prompt-Level Disagreement between Perceptual Quality and Alignment

Brian Rogers (IQ), Cody Walker (IL), Julian Hoffman (GD)

Abstract

Averaging generated-image scores by prompt changes the relationship observed between quality and alignment. In AGIQA-3K, the correlation between the 300 prompt-level means is 0.545, with a rank correlation of 0.529. The average of the ten within-condition correlations is 0.500. This secondary analysis shows that aggregation units carry substantive meaning. It does not identify a cause for the remaining disagreement or validate a new image-quality model. The results support keeping semantic and perceptual outcomes separate when evaluating prompts and generators.

1. The prompt is a different observation unit

A row-level correlation describes how individual released image scores move together in the observed table. A prompt-level correlation instead compares averages across the generation conditions available for each prompt. These quantities need not agree because they assign variation to different units. A strong relationship at one level cannot automatically be transported to another. This study asks what the public score record shows when the prompt, rather than the individual image row, becomes the descriptive comparison unit.

Decoupling Semantics from Distortions motivates a distinction between semantic and perceptual evidence in generated-image assessment (Meng, 2026). The present study uses that distinction to interpret human quality and alignment means. It does not compute the focal architecture's predictions or reproduce its multiscale streams. Its empirical contribution is a transparent aggregation analysis using released scores, with attention to which prompt-condition observations exist and how their averages define the resulting relationship.

AGIQA-3K supplies observations of human judgments about generated images (Li et al., 2023). Its released score table can support a secondary analysis without downloading the images, but that analysis has a defined information boundary. It does not observe new image features, rerun a quality model, or recover each participant's response from a mean and standard deviation. Dataset documentation should connect rows to prompts, generation conditions, collection decisions, and intended use (Gebru et al., 2021). Reproducible software records also matter when constructing splits or transforming variables (Pedregosa et al., 2011). The analyst should retain source labels rather than reinterpret a filename as a fully controlled experimental intervention. A prompt repeated across conditions creates useful pairing, yet the published table may still contain missing cells or unmeasured generation differences. A transparent report states which records are present, how conditions were identified, and which outcome was actually analyzed. Those details make a modest score-based result independently understandable without expanding it into an unperformed assessment of image content or model performance.

2. Constructing prompt summaries from observed rows

The released table contains 2,982 rows and 300 unique prompts across ten condition labels derived from filenames. For each prompt, the analysis takes the arithmetic mean of the observed quality means and the arithmetic mean of the observed alignment means. It also records the number of contributing rows. No missing prompt-condition score is imputed. The resulting summaries therefore describe available observations rather than an assumed complete generation experiment with every cell filled.

There are 287 prompts with ten rows, nine with nine rows, three with eight rows, and one with seven rows. These counts show that most prompt means have complete condition coverage while a small subset represents a different mixture. A complete-prompt analysis would define another descriptive population. The present calculation retains all 300 prompts and makes their unequal support explicit. It does not treat the missing rows as measured failures or infer why those combinations are absent from the source release.

Pearson correlation across prompt-level quality and alignment means is 0.545. Spearman rank correlation is 0.529. Separately, the analysis computes a Pearson correlation within each of the ten condition groups and takes their unweighted arithmetic mean, yielding 0.500. That latter summary gives each condition equal weight; it is not a pooled correlation and should not be interpreted as one. Neither statistic by itself measures the frequency of consequential disagreement on individual images.

Precision and identification are separate properties of a result. A narrow interval can surround an association whose scientific target is unclear. Robust covariance calculations allow particular departures from simple error assumptions, including unequal variance (White, 1980) and dependence within declared clusters (Cameron and Miller, 2015). Their role is to improve uncertainty calculations under those assumptions. They do not establish a causal assignment, guarantee a correct measurement, or expand coverage to conditions missing from the data. The analyst should therefore explain both the averaging population and the source of replication. An evaluation repeated many times on one history may provide information about stochastic variability while supplying little evidence about transfer to another history. Those two sources of uncertainty should remain distinguishable in the final interpretation.

An assessment instrument should be chosen for a stated target. CLIP learns visual-language correspondences (Radford et al., 2021), whereas MUSIQ addresses image quality using multiscale processing (Ke et al., 2021). NIMA's learned assessment setting (Talebi and Milanfar, 2017) and NIQE's natural-image-statistics approach (Mittal et al., 2013) introduce further distinctions in what evidence enters a score. Their relevance is not that one number can replace all the others. It is that perceptual appearance, aesthetics, and instruction correspondence need explicit definitions and validation populations. An image may rank differently under those requirements for meaningful reasons. A report should preserve the component outcomes before considering any combined utility. If only published human means are analyzed, the study can examine the relationship among those means but cannot claim a new benchmark result for an automated architecture. A further experiment would need actual predictions, appropriate splits, and independent outcome information. Keeping those requirements separate allows a score-based empirical result to remain useful without borrowing performance claims from methods that were never run.

An evaluation protocol should state what remains unfamiliar at assessment time. Leaving out an observation while retaining its close relatives can test interpolation rather than the broader transfer suggested by the study's wording. Validation methods for structured data emphasize matching the split to temporal, spatial, or hierarchical dependence (Roberts et al., 2017). Separately, parameter selection can bias an evaluation if the same observations guide both choices and reporting (Varma and Simon, 2006). These issues are not solved by calling a split a test set. The record should identify shared sources, the information accessible during development, and the stage at which each design choice was fixed. A result can be informative within a narrow boundary while providing no direct evidence about a more distant deployment population.

3. Reading associations at their stated level

Prompt-level averaging can reduce some image-specific variation while preserving differences in the conditions contributing to each prompt. It can also combine favorable and unfavorable outcomes that cancel in the mean. A prompt with similar average quality and alignment may still include individual images where the two judgments diverge sharply. The present result therefore describes relative prompt summaries, not universal agreement between perceptual quality and semantic success within every generated example.

The remaining disagreement should not be assigned a unique cause. It could relate to prompt content, generation behavior, assessment conventions, or unobserved characteristics of individual images. The table-based analysis has not processed image pixels or participant-level responses. An adjusted or experimental follow-up could investigate specified explanations, but a correlation below one does not identify which mechanism produced it. The useful observation is the coexistence of positive association and incomplete correspondence under a clear aggregation rule.

For model evaluation, prompts should remain grouped when they define the intended assessment boundary. A predictor assessed on new prompts faces a different task from one assessed on new images of familiar prompts. The present descriptive calculation is not a prediction experiment, yet it helps motivate that design choice. It also suggests reporting separate quality and alignment outcomes at the level relevant to the application. A single combined score would require an explicit preference rule and would conceal information about their observed disagreement unless the components remain available.

A metric's original evaluation setting matters when interpreting a generated-image result. KonIQ-10k (Hosu et al., 2019) and KADID-10k (Lin et al., 2019) offer different quality-assessment populations and distortion structures. Structural similarity (Wang et al., 2004), learned perceptual correspondence (Zhang et al., 2018), and feature-distribution comparison through FID (Heusel et al., 2017) provide further examples of distinct measurement units. Some compare paired images; another concerns a collection's distribution. Treating them as equivalent quality labels can obscure the information needed to compute and validate each. A generated-image study should identify whether it has an appropriate reference image, an independently sampled comparison distribution, or only released human judgments. The answer determines which empirical claims are available. A table-based regression may reveal a useful relationship between human outcomes while saying nothing directly about an uncomputed metric. A later experiment can assess such metrics under matched data and declared splits, but its proposed procedure should remain separate from the observations already present in the released score table.

Describing a contrast accurately can be more valuable than giving it an unjustified mechanistic name. A causal interpretation needs assumptions about how exposure and outcome arose, including selection into the analyzed sample. Causal diagrams provide a language for those assumptions (Greenland et al., 1999). If several alternative metrics or partitions are examined, the investigator should also account for the freedom to select whichever result appears strongest. False-discovery control is relevant to explicit collections of statistical tests (Benjamini and Hochberg, 1995), while a descriptive sensitivity analysis should be presented as such. A model can fit the observed rows well and still leave the principal explanation unresolved. Reporting that boundary preserves the value of the observation without overstating what adjustment or a significance threshold has established.

Replacing an expensive judgment with an automated score changes the measurement process. A positive correlation is evidence of co-movement, but it does not establish that the two procedures can be used interchangeably (Bland and Altman, 1986). A calibrated forecast has a different requirement: its stated probabilities should match relevant observed frequencies (Guo et al., 2017). Neither property follows

merely from a favorable overall ranking. A useful comparison preserves paired measurements, identifies the conditions in which disagreement matters, and keeps scale transformations explicit. If the intended judgment concerns recognizability, playability, or physical compatibility, the proxy should be assessed against that judgment directly. Agreement on easy cases cannot supply evidence about the difficult conditions absent from the comparison.

4. Measurement, transfer, and reporting

Human preference introduces another target beyond quality and prompt alignment. Pick-a-Pic supplies comparative preference data (Kirstain et al., 2023), and Human Preference Score investigates preference-oriented assessment (Wu et al., 2023). DINO's self-supervised visual representation (Caron et al., 2021) can support another view of image content, but its training objective does not define the user's complete utility. Reference-conditioning methods such as DreamBooth (Ruiz et al., 2022) and IP-Adapter (Ye et al., 2023) also change what information generation receives. A quality comparison should record that input boundary and avoid treating a reference-rich workflow as equivalent to one with less subject evidence. Preference, recognizability, prompt execution, and perceptual appearance can disagree. The useful scientific response is to retain the relevant outcomes and state any application-specific weighting. A released mean-score analysis can motivate these distinctions without collecting preferences or running conditioning models. Its contribution should remain the observed relationship among available judgments, while further claims about user choice and reference-driven generation require the corresponding independent measurements.

Coverage should be reported whenever the system or its evaluator can avoid difficult cases. An abstention policy changes the population on which an accuracy or fidelity score is calculated. Selective classification formalizes the joint reporting of coverage and conditional performance (Geifman and El-Yaniv, 2017). Distribution-shift research adds that estimated uncertainty can behave differently outside the development population (Ovadia et al., 2019). A benchmark should therefore retain which observations were eligible, accepted, rejected, or unavailable. A result on a carefully limited subset can be useful, provided the subset is visible. The challenge is to prevent a narrow success claim from becoming a broad capability claim through omission of the denominator or the circumstances in which the system declines to operate.

The unit of provenance should be fine enough to reconstruct an important disagreement. Reusable identifiers can connect a reported score to its input evidence and the relevant model and evaluator versions (Wilkinson et al., 2016). Pipeline dependencies should be treated as part of the experiment because apparently minor changes can alter output eligibility or metric behavior (Sculley et al., 2015). A report that preserves only aggregate numbers prevents this reconstruction. Keeping per-case results and documented transformations enables alternative weighting and error analysis without repeating expensive generation. Such a record also helps identify whether two papers compare the same scientific objects or merely use similar names for different releases. Versioned evidence makes progress easier to assess across successive evaluations.

5. Conclusion

Quality and alignment have a positive but incomplete relationship across prompt-level averages in the released AGIQA-3K table. The result depends on the aggregation unit and observed condition coverage. It neither validates an automated architecture nor explains why particular prompts diverge. Its contribution is a reproducible prompt-level description that keeps semantic and perceptual outcomes distinct and clarifies the additional measurements needed before their disagreement receives a mechanistic or predictive interpretation.

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.