Sampling Conditions and Paired Judgments in Released Midjourney Scores

Adrian Bryant (HU), Christian Crawford (GN), Owen Thompson (LT)

Abstract

The two released Midjourney condition labels in AGIQA-3K differ strongly in their paired human quality scores. Across 296 shared prompts, lowstep minus normal averages -1.239 quality units and -0.495 alignment units. Quality decreases in 287 pairs and increases in nine. These are descriptive comparisons of published rows, not a new intervention on a current generation service. The analysis makes the prompt pairing and incomplete table support explicit while keeping perceptual quality and instruction alignment as separate observed outcomes.

1. A condition label defines a bounded comparison

Public benchmark labels can preserve useful distinctions without documenting every operational detail of the system that produced their images. Their score differences should therefore be interpreted at the level the release supports. This study compares the recorded midjourney_lowstep and midjourney_normal rows in AGIQA-3K. It does not assume that the labels identify every setting in a current implementation, and it does not infer that repeating the named condition elsewhere would necessarily reproduce the same numerical difference.

Decoupling Semantics from Distortions motivates separate consideration of semantic alignment and perceptual quality in generated-image evaluation (Meng, 2026). The present study uses both released human outcomes rather than merging them into a single success label. It does not run the focal quality model or compare automated predictors. Its empirical contribution is a transparent paired description, showing which prompts contribute, how the average differences behave, and what remains unknown about the generation and rating process behind the source rows.

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.

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. Matching prompts and describing the contrast

The full score table contains 300 prompts and ten generation-condition labels. Each of the two Midjourney conditions has 296 observed rows. Their paired comparison uses the 296 shared prompts with available scores in both conditions. The difference is always lowstep minus normal. No missing score is replaced with a failure value, and no absent row is imputed. All calculations use the public means and preserve the prompt as the pairing unit, without processing the underlying images.

Mean quality difference is -1.239, with a standard deviation of 0.655 across the paired prompt differences. Nine differences are positive and 287 are negative. This sign pattern is more concentrated than a comparison in which a small negative mean results from nearly balanced improvements and declines. The standard deviation describes heterogeneity among prompt contrasts; it is not a standard error of the mean and does not reconstruct participant-level rating uncertainty. The present report does not manufacture a confidence interval from assumptions that have not been separately justified.

Mean alignment difference is -0.495 for the same paired comparison. Both outcomes therefore decline on average in the recorded lowstep condition, but their magnitudes are on their respective judgment scales. A larger numerical quality difference should not automatically be described as a proportionally larger practical effect than the alignment difference. That interpretation would require a justified relationship between scales and an application-specific consequence. The useful observation is that the two outcomes move together in average direction while retaining different numerical and conceptual meanings.

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.

Quality assessment is a family of measurement problems rather than a single numerical target. CLIP supplies a visual-language representation (Radford et al., 2021); MUSIQ investigates multiscale quality assessment (Ke et al., 2021); NIMA concerns learned image assessment (Talebi and Milanfar, 2017); and NIQE draws on natural-image statistics (Mittal et al., 2013). The different foundations motivate explicit separation of appearance and semantic correspondence in a generated-image study. A model may reproduce the requested content with visible defects, while another produces a polished but semantically incomplete image. Neither outcome is fully described by the other dimension. A useful report would identify its chosen endpoint, the evidence supplied to its assessor, and the conditions under which that interpretation has been checked. When the empirical input is a table of published human scores, the direct result concerns score relationships under that collection process. It cannot determine which automated representation caused those relationships or how a new predictor would generalize without additional predictions and a properly separated assessment.

3. Interpreting a strong descriptive pattern

The paired design holds prompt identity constant, which is an advantage over comparing unrelated prompt sets. It does not hold all other unrecorded factors constant. The released table is not a newly randomized repeated-sampling experiment. It does not provide the present analysis with complete generation histories or individual rater responses. A causal explanation about one sampling parameter would therefore require additional controlled evidence, even when the observed score difference is large and its sign is consistent across most prompts.

Missing table support should also remain visible. The nominal prompt universe contains four prompts absent from each of these condition groups. The paired estimate describes the observed shared prompts rather than every possible prompt or all users of a generation system. No conclusion about the absent outputs follows from their missing rows. A future completeness audit can investigate the source record, while a new collection can assess whether the descriptive contrast persists under a deliberately documented generation protocol.

For image-quality assessment, this comparison offers a demanding but bounded target. A future predictor could be tested on whether it captures the direction and heterogeneity of these human-score contrasts using appropriate independent inputs. That test would need actual predictions and a split that respects prompt dependence. The current result supplies neither. It motivates retaining paired quality and alignment evidence, so that a model's apparent success is not established merely by reproducing a broad condition ranking while overlooking important variation across individual prompts.

Quality evidence is shaped by both the dataset and the instrument. KonIQ-10k (Hosu et al., 2019) and KADID-10k (Lin et al., 2019) illustrate different collection and distortion settings. Structural similarity (Wang et al., 2004) and learned perceptual metrics (Zhang et al., 2018) compare image pairs under their own representations, whereas FID (Heusel et al., 2017) concerns feature distributions across collections. A report should preserve these distinctions when discussing generated imagery. An individual image score cannot establish an entire generator's distributional fidelity, and a collection metric cannot reveal every local semantic failure. Reference-based measures additionally depend on the suitability of the reference. These observations motivate a layered assessment rather than an unexplained total. A score-table study can examine how available human judgments co-vary and where their rankings disagree. It should not present that analysis as an empirical test of metrics for which no predictions were computed. The missing measurement can be specified in a future protocol without being silently filled by a citation to an established method.

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. From score associations to an assessment design

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 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.

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

The released Midjourney pair has a large negative average quality contrast and a negative alignment contrast, with quality declines in most observed prompt pairs. This is a strong descriptive pattern within a defined public table. It is not a causal estimate of a current service's sampling settings. Its contribution is a reproducible paired comparison that preserves outcome distinctions, source labels, and incomplete coverage instead of extending the available evidence beyond the recorded generation conditions.

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.