Paired Prompt Comparisons among Stable Diffusion Settings

Patrick Walsh (PL), Tyler Hughes (JO), Nicholas Bailey (AU)

Abstract

Pairing AGIQA-3K rows by prompt reveals small average quality differences but substantial prompt-to-prompt variation among four recorded Stable Diffusion conditions. Relative to sd1.5_normal, mean quality differences are -0.088 for highcorr, -0.145 for lowcorr, and -0.070 for lowstep. Each contrast contains 300 prompt pairs. The study describes released judgments and does not rerun generation settings. It illustrates why an average condition difference should be accompanied by its pairing structure and the distribution of individual prompt contrasts.

1. Pairing makes the comparison's unit explicit

When the same prompt appears under several conditions, a paired analysis can compare its released scores directly. This removes differences in which prompt identities enter a contrast, although it does not control every feature of image generation or rating. The present study uses that pairing to describe three condition differences within the public AGIQA-3K table. It asks whether their average differences conceal substantial variation across the prompts that contribute to each comparison.

The semantic and perceptual separation proposed in Decoupling Semantics from Distortions motivates reporting both human quality and alignment contrasts (Meng, 2026). The focal architecture is not reproduced here, and no new image-quality predictions are obtained. The empirical contribution is a paired secondary analysis of published means. It should not be interpreted as an implementation-level experiment on the named generation settings, whose labels alone do not supply every parameter, seed, or processing detail needed for that stronger claim.

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.

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.

2. Reconstructing the released prompt pairs

The input contains 300 rows for each of four source condition labels: sd1.5_normal, sd1.5_highcorr, sd1.5_lowcorr, and sd1.5_lowstep. For each outcome, the table is reshaped with prompt identifiers as rows and condition labels as columns. Differences subtract the normal-condition score from the alternative-condition score. All three quality contrasts have 300 observed pairs. No missing value is imputed and no new image is generated. The analysis retains the source labels without expanding them into undocumented experimental descriptions.

Mean quality difference is -0.088 for highcorr, -0.145 for lowcorr, and -0.070 for lowstep. Their corresponding standard deviations across prompt differences are 0.834, 0.739, and 0.768. These spreads are much larger than the absolute means, indicating considerable variation among the observed prompt contrasts. They are standard deviations of paired score differences, not standard errors and not estimates of individual rater disagreement. The unit is the prompt's difference between two released image-score rows.

The alignment differences also average below zero: -0.069 for highcorr, -0.146 for lowcorr, and -0.041 for lowstep. Quality improves for 144 prompts and declines for 155 in the highcorr comparison, with one tie. Lowcorr has 130 positive and 170 negative quality differences. Lowstep has 150 positive and 149 negative differences, again with one tie. Thus a negative average does not imply that every prompt, or even a large majority in every contrast, receives a lower quality score.

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.

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.

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.

3. What the paired pattern supports

The signs and spreads describe heterogeneity in the observed table. They do not identify which prompt property explains it. The analysis has not processed image content, parsed prompt semantics, or estimated effects of particular words. A follow-up could declare such features and test specified relationships, but exploratory explanations should not be assigned from the paired summary alone. The present result is useful because it shows that one overall mean is an incomplete description of the released condition contrast.

Pairing also does not establish a causal setting effect. The table offers one observed score row per prompt-condition combination, not a newly randomized sequence of repeated generation trials. Unrecorded variation in outputs, sampling, or assessment may contribute to each difference. A direct intervention study would need known implementations, controlled settings, repeated draws, and an appropriate rating protocol. Its results could be compared with these released-score observations, but the two evidence types should remain distinguishable in the final interpretation.

For model evaluation, the comparison suggests retaining quality and alignment differences at the prompt level rather than reporting only a pooled ranking. A predictor could match average condition differences while missing which prompts improve or decline. Testing that proposition would require predictions and a held-out design. The current study does not claim such a test. It provides a reproducible descriptive baseline and identifies a future assessment target: whether an instrument captures the observed heterogeneity without using the final human quality scores as input.

The phrase generalization needs an object: another observation of the same source, another source within the same collection, or a different acquisition process. Each requires a different assessment boundary. Structured validation makes explicit how correlated observations are assigned across training and evaluation (Roberts et al., 2017). Selection bias can still arise if the final boundary is repeatedly consulted while choosing a method (Varma and Simon, 2006). Publishing the split identifiers and the information permitted during each stage allows the claimed transfer to be checked. When a study proposes a new architecture but performs only a data audit, it should state that difference directly. Useful evidence about data coverage does not require pretending that a prediction system has already been trained and externally validated.

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. From score associations to an assessment design

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.

A system's reliable operating region may be narrower than the collection of inputs an interface accepts. Selective prediction separates performance on accepted cases from the fraction of cases accepted (Geifman and El-Yaniv, 2017). Reporting only the former can hide a substantial loss of coverage. Uncertainty estimates may also change under distribution shift (Ovadia et al., 2019), so familiar-condition confidence is not automatically evidence about new subjects, scenes, or control histories. A useful evaluation records rejected, ambiguous, and unsupported cases alongside successful ones. This makes it possible to distinguish limited but honest capability from a system that answers broadly with unreliable certainty. Any proposed deferral rule still requires assessment on its intended input population; it is not validated merely by being described.

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

The paired Stable Diffusion labels have modest negative average contrasts and substantial prompt-level variation in released quality judgments. Alignment differences are also negative on average, but individual quality signs are mixed. These findings describe the source table rather than a new settings experiment. Their contribution is a clear paired estimand and a warning against turning an average score difference into a universal statement about how every prompt responds to a generation condition.

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.