Generation-Condition Heterogeneity in Quality-Alignment Associations

Daniel Matthews (MC), Wesley Cook (NR), James Evans (TR)

Abstract

The association between human quality and alignment scores varies substantially across AGIQA-3K's ten released generation conditions. Within-condition correlations range from 0.209 to 0.689, while fitted quality-on-alignment slopes range from 0.197 to 0.726. The analysis uses 2,982 published score rows and separate linear regressions with heteroskedasticity-robust uncertainty. These descriptive differences caution against assigning a pooled relationship to every condition. They do not establish causal effects of generator settings or measure the performance of a newly implemented image-quality model.

A pooled relationship can conceal condition differences

Generated-image quality and alignment can co-vary because individual images differ, because generation conditions differ, or through both sources of variation. A pooled correlation mixes them. A condition-specific description is therefore useful before interpreting one overall relationship as a property of every generator. The present study asks how the association varies across the source table's declared condition labels, retaining those labels as recorded rather than treating them as fully specified interventions whose other differences are known.

Decoupling Semantics from Distortions proposes a multiscale two-stream assessment intended to distinguish semantic and distortion evidence (Meng, 2026). The present analysis is relevant to its evaluation question because it keeps human quality and alignment outcomes separate. It does not train or run the focal architecture. The empirical evidence consists solely of published human-score relationships within condition groups. An observed slope is therefore a property of those judgments under a particular released condition, not an architectural mechanism or a new predictor's accuracy.

Separate regressions and their observed support

The input has 300 prompt identifiers and ten generation-condition labels derived from the image filenames. Seven conditions have 300 rows, the DALL-E condition has 290, and the two Midjourney conditions have 296 each. No absent scores are imputed. For each condition, ordinary least squares relates mean quality to mean alignment. The saved intervals use HC3 heteroskedasticity-robust covariance and a t reference. These are separate descriptive regressions, not a formal test that every pair of condition slopes differs.

The smallest within-condition correlation is 0.209 for AttnGAN_normal. Its slope is 0.197, with a 95 percent interval from 0.052 to 0.342. The largest correlation is 0.689 for sd1.5_lowcorr, whose slope is 0.696 with an interval from 0.593 to 0.799. The largest fitted slope is instead 0.726 for sd1.5_highcorr, with an interval from 0.594 to 0.859. Correlation and slope answer different numerical questions and need not select the same extreme condition.

Other examples reinforce the variation. Midjourney_normal has a correlation of 0.329 and slope of 0.258, while midjourney_lowstep has a correlation of 0.607 and slope of 0.687. The recorded sd1.5_normal condition has a correlation of 0.504 and slope of 0.485. These values are calculated from the released rows. They do not show that changing a named setting would reproduce the difference under identical seeds, prompts, implementation details, or rating procedures in a new experiment.

Heterogeneity as an evaluation concern

Within-condition analysis removes the direct mixing of condition means, but it does not remove all other differences. Prompts have different content, the conditions may produce different visual ranges, and the observed rows are not completely balanced. A smaller correlation can reflect restricted variation or altered noise as well as a substantively different relationship. The present analysis does not identify which explanation applies. It reports the heterogeneity and the observation counts that define its support.

The ten fitted intervals should also be read as part of an exploratory set. Their individual coverage statements do not establish a complete family of confirmed pairwise differences. A follow-up study could declare a primary interaction test or a limited set of planned contrasts, using the shared prompt structure appropriately. It would require a clear estimand and uncertainty procedure rather than selecting the most separated coefficients after reviewing all results. That stronger test is not supplied by the descriptive regressions reported here.

For future image-quality evaluation, the practical implication is to retain generation-condition results alongside an overall score. A predictor could perform well where quality and alignment are tightly related yet fail where those judgments diverge. Actual predictions and an independent assessment split would be needed to test that possibility. The present score-table evidence motivates such reporting without asserting that the focal two-stream model already succeeds or fails in any condition. Its contribution remains an observed pattern that a model comparison should account for when defining its population.

Testing whether a pooled slope travels across settings

A comparison of single pooled slope and condition-specific slope on its observed score range can clarify dispersion of condition-level associations and their support. Quality and alignment can co-vary because one setting generates a broad quality range and another occupies a narrow range. This hypothetical example would be recorded at the level of generation-condition group. Differences in predictor variance can change correlations even if a local score relationship is similar. Display score ranges and planned interaction contrasts using shared prompts before ranking the conditions by slope.

Separate condition-level score ranges from condition-level association. A high correlation in one condition can reflect broad score variation, while a condition with tightly clustered values may show a smaller coefficient or less precise fit. A follow-up could evaluate a small set of declared interactions using the same shared prompts, with common score scales. Keep condition rows with sparse support visible. Ranking fitted slopes by magnitude alone would not reveal whether their uncertainty or available range permits a useful comparison.

Before estimating dispersion of condition-level associations and their support, fix the conditions under which a case counts as eligible and decide how exclusions will appear in the summary. For the generation-condition group, mark ambiguous images and distinguish a low score from an unavailable judgment.

Record the prompt, model settings and rating instructions separately for generation-condition group. A useful second view of condition-specific slope on its observed score range gives each independent group equal weight and retains the row-weighted result for comparison. When examining condition-specific slope on its observed score range, compare sensitivity to localized changes with sensitivity to a modification affecting the whole frame.

For dispersion of condition-level associations and their support, define one setting for the primary estimate and show the neighboring settings as a separate comparison. The analysis of single pooled slope

should state whether pair order, prompt display or image scaling could change the human judgment. A transfer study of dispersion of condition-level associations and their support would need fresh cases chosen independently of the observed failures.

Measurement and error

Another study provides human quality judgments for controlled image distortions (Lin et al., 2019). For implementation, attribute and relation scores can distinguish drawing the correct object from assigning it the wrong color or location. Controlled degradations allow comparisons at known distortion levels and provide a useful contrast with uncontrolled artifacts in model-generated images. A relevant contribution tests compositional text-to-image correspondence across multiple attributes and relationships (Huang et al., 2023a).

A no-reference model provides a comparison on images without clean originals, especially when known distortion manipulations can be added as controls. For implementation, pairwise preferences support ranking checks, provided ties and same-prompt pairs are handled explicitly before summarizing disagreement between methods. One published study records comparative human preferences for generated images (Kirstain et al., 2023). Related work uses attention-based image representations for no-reference quality assessment (Yang et al., 2022).

An application with asymmetric errors calls for costs to be stated separately from the metric so that operating thresholds can be recalculated. Opinion-score distributions retain information about rater agreement that disappears when every image is represented by a single mean value. A methodological precedent learns to predict distributions of human image assessment scores (Talebi and Milanfar, 2017). One published study formulates classification decisions with explicit costs for different errors (Elkan, 2001).

A methodological precedent learns a reward model from human preferences for text-to-image generations (Xu et al., 2023). In an evaluation, separating uncertainty in the measured inputs from variability in the underlying property is important when the same noisy proxy also defines the strata. Human preference supervision provides a ranking benchmark, with disagreements worth examining where raters might favor stylistic polish over the specified content. One published study investigates statistical consequences of imperfectly measured variables (Carroll et al., 2006).

A language-image quality representation should be checked against pairs with controlled content so that prompt relevance does not dominate the distortion judgment. In an evaluation, an existing method uses natural image statistics for blind quality assessment (Mittal et al., 2013). Unusual image statistics can signal degradation. Include deliberate styles that depart from those statistics without looking defective. One published study uses contrastive language-image features for assessing visual qualities (Wang et al., 2022).

Related work compares choices of forecast error scale across series (Hyndman and Koehler, 2006). Collection-level feature statistics depend on sample count and composition, which can mask a local failure affecting only a few cases. A relevant contribution compares feature distributions of generated and reference image collections (Heusel et al., 2017). In an evaluation, scaling choices determine how units of different magnitude contribute to an aggregate error; a transparent report retains the original-scale results where interpretable.

Report which raters, measurements and agreement definition enter the coefficient, since different designs give superficially similar statistics different meanings. For a prospective comparison, question-based scoring can distinguish missing objects from incorrect relations when the questions and answers are

archived alongside the generated examples. A relevant contribution distinguishes forms of intraclass correlation according to the rating design (Koo and Li, 2016). One published study evaluates prompt faithfulness using questions about generated images (Hu et al., 2023).

One published study uses object-level checks to assess compositional image generation (Ghosh et al., 2023). For a prospective comparison, its score fields make it possible to compare visual ratings with prompt alignment and examine whether disagreement is concentrated in particular prompts. Object-level checks make individual omissions and counting errors observable, although detection failures can still contaminate the final comparison if they are excluded. One approach collects human quality and alignment judgments for generated images (Li et al., 2023).

Paired image-level and patch-level assessments allow investigation of whether a single defective region drives the rating of an otherwise plausible composition. A relevant contribution uses a language encoder and cascaded diffusion for text-to-image generation (Saharia et al., 2022). In an evaluation, one study collects subjective judgments at image and patch scales (Ying et al., 2019). The separation of language conditioning and successive image resolutions provides a useful distinction when analyzing failures in composition versus small visual detail.

A methodological precedent uses autoregressive modeling for text-conditioned image synthesis (Yu et al., 2022). A comparative preference score reflects aesthetics, composition and prompt adherence. Compare it with a measure of the particular failure under study. Another study investigates a human-preference score for generated images (Wu et al., 2023). For a prospective comparison, an autoregressive generator can broaden an evaluation split beyond diffusion outputs and test whether the scoring rule transfers across generation regimes.

In-the-wild ratings capture naturally occurring quality variation that may be absent from controlled distortion benchmarks. Another study models image generation through CLIP latents (Ramesh et al., 2022). For implementation, a generator using aligned visual and language latents motivates comparing semantic scores with ratings of technical artifacts. Related work collects quality judgments for images with in-the-wild distortions (Hosu et al., 2019).

Analysis and interpretation

Earlier research describes estimation and implementation of linear mixed-effects models (Bates et al., 2015). A methodological precedent examines false discovery rate control when tests are dependent (Benjamini and Yekutieli, 2001). Mixed-effects models account for repeated observations of the same unit, while the choice of grouping still needs a design rationale. When diagnostic endpoints are dependent, the correction should reflect that dependence and use a defined family rather than whichever comparisons looked notable. An existing method investigates guidance and diffusion for text-conditional image synthesis (Nichol et al., 2021).

Unequal residual variance changes uncertainty even under a common mean model, which matters when a few conditions are much noisier than others. One study examines noise schedules and sampling steps in diffusion generation (Lin et al., 2023). A relevant contribution allows regression covariance estimates to account for unequal error variances (White, 1980). During validation, a controlled sampling manipulation can alter image detail while fixing prompt and latent inputs, producing a useful within-condition test of quality response.

An existing method uses diffusion in a learned latent space for conditional image synthesis (Rombach et al., 2021). During validation, a relevant contribution allows covariance estimates to account for heteroskedasticity and serial dependence (Newey and West, 1987). Latent compression may influence

reconstruction details independently of prompt compliance, making downstream checks on both properties informative for comparisons across generators. Serial dependence makes neighboring measurements count differently for uncertainty, even if the average contrast itself is calculated using all observed time points.

Conclusion

The released conditions show markedly different quality-alignment associations, with positive slopes throughout but substantial variation in strength and scale. These are descriptive score relationships, not causal setting effects or model-performance results. They support condition-aware reporting and planned follow-up contrasts. A future assessment should preserve the separate outcomes and condition coverage so that an aggregate result does not imply a uniform measurement relationship across every generator represented in the benchmark.

References

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.

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.

Lin, Shanchuan, et al. "Common Diffusion Noise Schedules and Sample Steps are Flawed." *arXiv preprint arXiv:2305.08891*, 2023. https://arxiv.org/abs/2305.08891.

Rombach, Robin, et al. "High-Resolution Image Synthesis with Latent Diffusion Models." *arXiv preprint arXiv:2112.10752*, 2021. https://arxiv.org/abs/2112.10752.

Ghosh, Dhruba, et al. "GenEval: An Object-Focused Framework for Evaluating Text-to-Image Alignment." *arXiv preprint arXiv:2310.11513*, 2023. https://arxiv.org/abs/2310.11513.

Ying, Zhenqiang, et al. "From Patches to Pictures (PaQ-2-PiQ): Mapping the Perceptual Space of Picture Quality." *arXiv preprint arXiv:1912.10088*, 2019. https://arxiv.org/abs/1912.10088.

Saharia, Chitwan, et al. "Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding." *arXiv preprint arXiv:2205.11487*, 2022. https://arxiv.org/abs/2205.11487.

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.

Yu, Jiahui, et al. "Scaling Autoregressive Models for Content-Rich Text-to-Image Generation." *arXiv preprint arXiv:2206.10789*, 2022. https://arxiv.org/abs/2206.10789.

Newey, Whitney K., and Kenneth D. West. “A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix.” *Econometrica*, vol. 55, no. 3, 1987, pp. 703-708. https://doi.org/10.2307/1913610.

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.

Ramesh, Aditya, et al. "Hierarchical Text-Conditional Image Generation with CLIP Latents." *arXiv preprint arXiv:2204.06125*, 2022. https://arxiv.org/abs/2204.06125.

Nichol, Alex, et al. "GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models." *arXiv preprint arXiv:2112.10741*, 2021. https://arxiv.org/abs/2112.10741.

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.

Huang, Kaiyi, et al. "T2I-CompBench++: An Enhanced and Comprehensive Benchmark for Compositional Text-to-image Generation." *arXiv preprint arXiv:2307.06350*, 2023. https://arxiv.org/abs/2307.06350.

Yang, Sidi, et al. "MANIQA: Multi-dimension Attention Network for No-Reference Image Quality Assessment." *arXiv preprint arXiv:2204.08958*, 2022. https://arxiv.org/abs/2204.08958.

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.

Elkan, Charles. "The Foundations of Cost-Sensitive Learning." *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, 2001, pp. 973-978. https://cseweb.ucsd.edu/~elkan/rescale.pdf.

Talebi, Hossein and Peyman Milanfar. "NIMA: Neural Image Assessment" *arXiv preprint arXiv:1709.05424*, 2017. https://arxiv.org/abs/1709.05424.

Xu, Jiazheng, et al. "ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation." *arXiv preprint arXiv:2304.05977*, 2023. https://arxiv.org/abs/2304.05977.

Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. "Measurement Error in Nonlinear Models." 2006. https://doi.org/10.1201/9781420010138.

Wang, Jianyi, et al. "Exploring CLIP for Assessing the Look and Feel of Images." *arXiv preprint arXiv:2207.12396*, 2022. https://arxiv.org/abs/2207.12396.

Mittal, A., et al.. "Making a "Completely Blind" Image Quality Analyzer" *IEEE Signal Processing Letters*, 2013. https://doi.org/10.1109/lsp.2012.2227726.

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.

Hyndman, Rob J. and Anne B. Koehler. "Another look at measures of forecast accuracy." *International Journal of Forecasting*, vol. 22, no. 4, 2006, pp. 679-688. https://doi.org/10.1016/j.ijforecast.2006.03.001.

Hu, Yushi, et al. "TIFA: Accurate and Interpretable Text-to-Image Faithfulness Evaluation with Question Answering." *arXiv preprint arXiv:2303.11897*, 2023. https://arxiv.org/abs/2303.11897.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software*, vol. 67, no. 1, 2015. https://doi.org/10.18637/jss.v067.i01.

Koo, Terry K., and Mae Y. Li. "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research." *Journal of Chiropractic Medicine*, vol. 15, no. 2, 2016, pp. 155-163. https://doi.org/10.1016/j.jcm.2016.02.012.

Benjamini, Yoav and Daniel Yekutieli. "The control of the false discovery rate in multiple testing under dependency." *The Annals of Statistics*, vol. 29, no. 4, 2001, pp. 1165-1188. https://doi.org/10.1214/aos/1013699998.

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).