An Incomplete Condition-Prompt Grid in a Generated-Image Benchmark

Patrick Sullivan (IN), Oliver Douglas (ZM), Patrick Porter (MY)

Abstract

AGIQA-3K's released score table contains 2,982 of the 3,000 cells in its nominal ten-condition, 300-prompt grid. Eighteen combinations are absent: ten under DALLE2_normal and four under each of the two Midjourney labels. Thirteen prompts lack at least one condition. This secondary audit identifies the denominator behind pooled and paired comparisons. It does not infer why rows are absent or assign outcomes to them. The findings make completeness a visible part of evaluating semantic and perceptual judgments rather than an implicit assumption.

1. Completeness belongs to the comparison definition

A table can contain no missing values in its observed score columns and still omit entire combinations of its design variables. These are different forms of incompleteness. An analysis that checks only whether each existing row has a quality mean may miss the absent prompt-condition cells. The present study reconstructs the nominal grid in the released AGIQA-3K score table and identifies which comparisons receive complete support from the rows actually available.

Decoupling Semantics from Distortions motivates evaluating semantic and perceptual information separately (Meng, 2026). That distinction also requires clarity about which observations contribute to each evaluation. The present study does not reproduce the focal model or calculate its predictions. It audits the public score table's support. Its contribution is a bounded data-quality finding: the location of absent combinations, not a diagnosis of generation failure, a claim about missingness mechanisms, or evidence about the unobserved quality of images not represented in the release.

A public benchmark table is a useful empirical object in its own right. AGIQA-3K records quality and alignment judgments for generated imagery (Li et al., 2023). Studying those released scores differs from generating new images or reproducing an image-quality architecture. The distinction should remain visible in the data description and conclusion. Dataset documentation provides a way to explain the collection, row identities, and intended scientific use (Gebru et al., 2021), while software documentation helps make processing and validation choices recoverable (Pedregosa et al., 2011). In particular, prompt identifiers should remain linked across generation conditions rather than disappearing in a flattened analysis. Missing prompt-condition combinations also need an explicit record. A complete-case contrast and an all-observed-row contrast can describe different populations even when both are calculated correctly. The relevant empirical contribution is therefore a reproducible relationship among the available judgments, with its selection and measurement limits stated. It is not evidence that any newly proposed predictor has been trained on images or validated against independent raters when those steps have not occurred.

2. Reconstructing the nominal grid

The input has 300 distinct prompt strings and ten condition labels derived by removing the final image index from source filenames. Their Cartesian grid has 3,000 possible cells. The observed table has 2,982 rows, with each observed prompt-condition combination unique. Reshaping quality scores into this grid reveals 18 absent cells. The audit uses the published metadata and means; it does not download or inspect visual files to determine whether an omitted image exists elsewhere or why its score is unavailable.

DALLE2_normal contributes ten absent cells. Midjourney_lowstep and midjourney_normal contribute four each. The other seven conditions have complete 300-row coverage in the released table. At the prompt level, 287 prompts have all ten conditions and thirteen have at least one absence. Nine prompts have nine observed rows, three have eight, and one has seven. The row-count distribution and missing-cell total are consistent descriptions of the same incomplete grid under different units of summary.

No outcome is imputed. An absent cell is not labeled a failed generation, a zero-quality result, or a randomly missing observation. Those interpretations would require evidence beyond the grid itself. The audit also distinguishes source-label completeness from a complete account of the generation process. A table with every nominal cell present could still omit repeated attempts or unrecorded settings. The present check establishes only the support of the released prompt-condition arrangement used by subsequent score analyses.

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.

Visual-language alignment and image quality have related but distinguishable measurement foundations. CLIP connects images and text through language-supervised representation learning (Radford et al., 2021). Its similarities can inform an assessment, but they do not automatically define perceptual quality or an application's complete notion of semantic correctness. MUSIQ develops a multiscale image-quality approach (Ke et al., 2021), NIMA studies learned aesthetic assessment (Talebi and Milanfar, 2017), and NIQE offers a natural-scene-statistics-based quality measure (Mittal et al., 2013). These methods illustrate different targets and information assumptions rather than interchangeable instruments. A generated image can look convincing while missing the instruction, or satisfy a prompt while containing visible artifacts. The evaluation should retain those possibilities instead of treating every score as a noisy observation of one universal quality. In a secondary analysis of released human judgments, relationships between quality and alignment describe those judgments. They do not establish how one of these automated instruments would perform unless its predictions are actually obtained and assessed under a suitable independent protocol.

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. Consequences for pooled and paired results

A pooled all-row average weights conditions according to their observed row counts. An equal-condition average uses another weighting rule. A prompt average over available rows can likewise combine a different condition mixture for incomplete prompts. These are legitimate descriptive quantities when named clearly, but they should not be presented as interchangeable summaries of a fully balanced experiment. The reconstructed grid allows a reader to determine which denominator and weighting population each calculation actually represents.

Paired comparisons need their own observed intersection. A contrast between complete conditions can retain all 300 prompts, while a contrast involving an incomplete condition may use fewer. Restricting every analysis to the 287 complete prompts would restore a common observed condition grid but change the prompt population. It would not explain the missingness or establish unbiased inference about omitted cells. A useful sensitivity report could present both estimands while keeping their support and interpretation distinct.

The pattern of absence alone cannot determine a missing-data mechanism. It does not establish independence from quality, prompt content, or generation difficulty. Those possibilities can be stated as research questions, but assigning one would exceed the available record. A future source audit or new controlled collection could add the required evidence. Until then, the practical response is explicit coverage and bounded conclusions: compare observed judgments transparently and avoid expanding a balanced-looking table format into an unsupported assumption that every intended outcome was measured.

Established quality datasets and metrics remind the analyst that observation populations differ. KonIQ-10k studies quality in a large in-the-wild image collection (Hosu et al., 2019), while KADID-10k supports assessment of declared image distortions (Lin et al., 2019). These are not automatically the same population as generated imagery. Structural similarity (Wang et al., 2004) and learned perceptual similarity (Zhang et al., 2018) also require an appropriate interpretation of the compared images. FID instead compares distributions in a learned representation (Heusel et al., 2017). None of these objects is interchangeable with an individual human alignment judgment. A study should state which references or populations a metric needs and whether those inputs exist in the experiment. Distributional quality cannot be inferred solely from one image's mean score, and a high pairwise resemblance can reward repetition rather than successful instruction execution. The relevant lesson for a score-only analysis is to keep its outcome boundary precise. It can describe published judgments and propose a future metric comparison without reporting that the comparison has already been performed.

An association can suggest a research direction without identifying an intervention. Causal diagrams make it possible to state which variables precede the proposed change, which are consequences, and which selection processes remain uncontrolled (Greenland et al., 1999). Their value lies in exposing assumptions rather than decorating a regression with a causal label. A second issue arises when many outcomes or subgroups are explored: selecting the most favorable comparison can exaggerate the impression of support. False-discovery procedures address a defined family of inferential tests (Benjamini and Hochberg, 1995). They do not turn an exploratory pattern into a replicated mechanism. A careful report distinguishes prespecified comparisons, descriptive sensitivities, and hypotheses requiring new observations. This distinction is especially important when an evaluation offers many plausible metrics.

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. From score associations to an assessment design

An image that receives a high quality judgment need not win every user comparison. Pick-a-Pic (Kirstain et al., 2023) and Human Preference Score (Wu et al., 2023) motivate attention to preference as a distinct assessed outcome. DINO (Caron et al., 2021) provides self-supervised visual features that can characterize content under another representation. DreamBooth (Ruiz et al., 2022) and IP-Adapter (Ye et al., 2023) illustrate how subject or image references can influence the generation task before scoring begins. These sources help explain why an evaluation should specify both its input evidence and its target judgment. A method with additional references or adaptation opportunities may solve a different information problem. Likewise, a predictor of aggregate alignment may require human information unavailable in a practical no-reference quality system. The report should make such access explicit. A secondary score-table analysis remains informative when it describes its available outcomes precisely and reserves statements about preference prediction, subject conditioning, or visual-feature performance for studies in which those outputs were actually obtained and independently assessed.

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.

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 benchmark has eighteen absent prompt-condition cells concentrated in three condition labels. This is a reproducible support finding, not an explanation of absence or an estimate of missing outcomes. It clarifies the populations behind pooled, prompt-level, and paired analyses. Future semantic and perceptual assessment should preserve this coverage information so that a score difference refers to the observations actually available rather than to an assumed complete experiment.

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.