Mean-Median Disagreement across Atari Task Distributions

Caleb Collins (PH), William Spencer (NL), Joseph Collins (DJ)

Abstract

Within-game means and medians often differ less than a benchmark's overall mean and median might suggest. In the public BBF files, their game rankings have Spearman associations of 0.995 and 0.982. Yet individual mean-minus-median gaps reach 0.870 in the main configuration and -0.589 in the opposite direction. These findings distinguish rank stability from agreement in magnitude. A secondary analysis of repeated scores motivates retaining both location summaries in interactive evaluations, especially when occasional strong or weak outputs matter differently to users.

1. Two uses of the word typical

The mean describes an expectation under a specified observation distribution. The median describes its middle location. Either can be a defensible account of a typical outcome, but they answer different questions when a distribution is asymmetric. A model that occasionally achieves very high scores can have a mean that few individual runs resemble. A model with occasional severe failures can have a favorable median while providing a less favorable expected experience. Choosing between the summaries therefore requires a purpose rather than a universal preference for one statistic.

Make a game concerns generated interaction, where a typical frame, a typical trajectory, and an expected user experience are different objects (Meng et al., 2026a). The source's interactive focus makes those distinctions worth preserving, although no output from that renderer is measured here. The empirical contribution instead examines means and medians within public BBF task distributions. This supplies a limited example of how a stable ranking can coexist with disagreement in the numerical outcomes being summarized.

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.

Common evaluation platforms make comparisons convenient, but convenience should not determine the scientific estimand. ALE organizes many distinct games under a reusable platform (Bellemare et al., 2012), while Gym supplies a widely used interaction abstraction (Brockman et al., 2016). Procgen emphasizes that procedural instances can themselves define an important assessment boundary (Cobbe et al., 2019). BBF's public result tables give a concrete example in which tasks and runs can be counted separately (Schwarzer et al., 2023). The broader statistical lesson is that aggregate performance depends on uncertainty, task heterogeneity, and the chosen summary (Agarwal et al., 2021). A rendering evaluation should therefore retain environment-level outcomes before constructing a total. Matching the number of generated frames does not necessarily match the diversity of actions or histories. Nor does a large number of repeated observations make the benchmark representative of every interactive setting. The relevant claim is determined by which tasks and variations were observed, under what controls, and with what weighting.

2. Computing the comparison at the game level

The two released score files include 55 games and 2,912 records in total. For each configuration-game group, the analysis calculates the mean and median of the supplied normalized score. Their difference is retained in the same normalized units. The operation is performed separately within each game, so it does not confuse variation across game baselines with within-game asymmetry. The group summaries are then described across the 55 tasks with equal task influence.

For the main file, the average mean-minus-median gap is 0.0479 and its median is 0.0008. For the file without resets, those values are 0.0093 and 0.0015. The near-zero median gaps show that a substantial discrepancy is not universal. Nevertheless, the main-file range extends from -0.5893 to 0.8699. The comparison-file range is -0.2255 to 0.2927. A concise average of gaps would therefore conceal cases with quite different distributional shapes and potential practical consequences.

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.

Learned world models connect visual representation to behavior in several ways. World Models explores compressed representations of environment experience (Ha and Schmidhuber, 2018). Dreamer uses latent imagination to support behavior learning (Hafner et al., 2019), IRIS studies transformer-based world modeling (Micheli et al., 2022), and DIAMOND investigates diffusion-based representations in which visual detail can matter to control (Alonso et al., 2024). These works establish a useful set of questions rather than a single interchangeable implementation. What information is retained, which histories condition prediction, and how is a predicted observation connected to an action consequence? A rendering study should state which of these questions it actually tests. Visually plausible outputs alone do not establish an accurate transition model. Conversely, an agent can obtain a favorable return without reproducing all details relevant to a human viewer. The measurement design should preserve this distinction and avoid translating a result from one endpoint into an unobserved improvement on another.

3. Rank agreement does not settle the choice

The mean exceeds the median in thirty main-file games and thirty-one comparison-file games. The sign counts alone do not reveal the severity or frequency of unusual outcomes. A positive gap can arise through a few very high scores or a broad moderate asymmetry. A negative gap can arise through an adverse lower tail. The available summary file supports describing the gap, while the underlying repeated scores are needed to inspect which pattern occurs in a particular game.

Across games, mean and median ranks are strongly associated: Spearman correlation is 0.9949 in the main file and 0.9816 in the comparison. This is evidence of broadly stable task ordering under the two location summaries. It is not evidence that their numerical values agree for every task. Ranking can remain almost unchanged when a small set of outcomes shifts the magnitude of several game means. The distinction matters if an evaluation uses a score difference to justify a resource allocation, a reliability claim, or a threshold rather than merely an ordinal comparison.

For rendering, a study could retain a distribution over repeated generations of a fixed history. A central summary would describe the usual experience, while tail-sensitive summaries would document rare but consequential control or visual failures. The appropriate endpoint should follow the use. An entertainment interface might tolerate a visually unusual but valid continuation, whereas a task requiring a persistent object may not tolerate a rare identity or state substitution. Neither a mean nor a median can supply that consequence without a definition of what constitutes an error.

An additional level is the distribution over histories. Taking a median across all output frames is not equivalent to taking a median within each trajectory and then averaging trajectories. Long clips and easy histories can receive different weight under those operations. The BBF example suggests reporting the hierarchy rather than flattening it. A later reader should be able to distinguish an average of task expectations from a statistic describing the middle task, because those summaries answer different population questions even when their rankings appear similar.

Generative interaction introduces measurement requirements beyond ordinary video synthesis. GameNGen studies neural generation supporting real-time game interaction (Valevski et al., 2024), while Genie investigates action-controllable environments learned from unlabeled videos (Bruce et al., 2024). Their objectives help distinguish response to control from the production of plausible motion. Structural similarity evaluates a defined image comparison (Wang et al., 2004), learned perceptual similarity uses a different visual representation (Zhang et al., 2018), and video-distribution metrics add temporal features to generative assessment (Unterthiner et al., 2018). None of these scores automatically establishes that an action had its intended effect. A practical study should preserve the action, its timing, and the relevant output history alongside visual measurements. It should also record cases where the system produces attractive imagery but violates a state or control requirement. Such disagreement can identify an important limitation that a combined average would hide. A multi-endpoint report is therefore more informative than an unexplained total labeled interactive quality.

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 useful reliability claim contains both a performance statement and a scope statement. Selective prediction shows why a system that answers only some cases must report how many it accepts (Geifman and El-Yaniv, 2017). Confidence estimation under a changed data distribution requires independent checking (Ovadia et al., 2019). These requirements apply to generators and to the automated judges used to assess them. If a judge cannot evaluate a profile, an occluded object, or an unusual interaction, that inability should remain visible rather than removing the case from the generator's score silently. Recording the entire decision path permits an analyst to identify whether an apparent improvement came from better outputs, narrower eligibility, or a more restrictive measurement procedure.

4. Interpreting playability evidence

The distribution of available behavior constrains what can be inferred from recorded interaction. The reinforcement-learning framework distinguishes the information in a trajectory from the policy that produced it (Sutton and Barto, 2018). Offline reinforcement learning makes the limitations of fixed experience explicit (Levine et al., 2020), and conservative Q-learning studies one way to address out-of-distribution action evaluation (Kumar et al., 2020). The corresponding lesson for rendering is methodological: an archive of familiar actions is not a complete intervention experiment. If commands are chosen by an existing policy, some states and responses may remain poorly represented. An evaluation should document this support and avoid treating absent cases as successful extrapolation. Proposed counterfactual tests should compare matched initial histories where feasible and retain whether a valid response was produced. Such evidence can strengthen a limited claim without implying that the system models every possible action or game mechanic.

An automated measurement should not inherit the authority of its training labels without a direct assessment of use. The relevant checks include systematic disagreement with the intended procedure (Bland and Altman, 1986) and calibration of any probability claims (Guo et al., 2017). Ranking performance, absolute accuracy, and confidence reliability are distinct outcomes. A study can improve one while leaving the others unchanged. Keeping the measurements separate permits examination of those tradeoffs and avoids implying that one favorable statistic establishes all aspects of validity. When independent judgments are unavailable, the limitation should remain part of the result. The absence of a target observation cannot be repaired by combining several correlated proxies into a more elaborate total score.

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 released Atari results show high rank agreement between within-game means and medians alongside some sizable task-specific magnitude gaps. This combination supports careful reporting rather than a blanket rejection of either summary. Interactive rendering studies should specify what typical means for the intended user and preserve the outcome distribution needed to assess rare consequences. A stable ranking is useful information, but it cannot establish numerical agreement or determine how occasional failures should be valued.

References

Meng, Zijie, et al. "Make a game: A novel paradigm for interactive game rendering." *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 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.

Bellemare, Marc G., et al.. "The Arcade Learning Environment: An Evaluation Platform for General Agents" *arXiv preprint arXiv:1207.4708*, 2012. https://arxiv.org/abs/1207.4708.

Agarwal, Rishabh, et al.. "Deep Reinforcement Learning at the Edge of the Statistical Precipice" *arXiv preprint arXiv:2108.13264*, 2021. https://arxiv.org/abs/2108.13264.

Cobbe, Karl, et al.. "Leveraging Procedural Generation to Benchmark Reinforcement Learning" *arXiv preprint arXiv:1912.01588*, 2019. https://arxiv.org/abs/1912.01588.

Brockman, Greg, et al.. "OpenAI Gym" *arXiv preprint arXiv:1606.01540*, 2016. https://arxiv.org/abs/1606.01540.

Ha, David and Jürgen Schmidhuber. "World Models" *arXiv preprint arXiv:1803.10122*, 2018. https://arxiv.org/abs/1803.10122.

Hafner, Danijar, et al.. "Dream to Control: Learning Behaviors by Latent Imagination" *arXiv preprint arXiv:1912.01603*, 2019. https://arxiv.org/abs/1912.01603.

Micheli, Vincent, et al.. "Transformers are Sample-Efficient World Models" *arXiv preprint arXiv:2209.00588*, 2022. https://arxiv.org/abs/2209.00588.

Alonso, Eloi, et al.. "Diffusion for World Modeling: Visual Details Matter in Atari" *arXiv preprint arXiv:2405.12399*, 2024. https://arxiv.org/abs/2405.12399.

Valevski, Dani, et al.. "Diffusion Models Are Real-Time Game Engines" *arXiv preprint arXiv:2408.14837*, 2024. https://arxiv.org/abs/2408.14837.

Bruce, Jake, et al.. "Genie: Generative Interactive Environments" *arXiv preprint arXiv:2402.15391*, 2024. https://arxiv.org/abs/2402.15391.

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.

Unterthiner, Thomas, et al.. "Towards Accurate Generative Models of Video: A New Metric & Challenges" *arXiv preprint arXiv:1812.01717*, 2018. https://arxiv.org/abs/1812.01717.

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.

Schwarzer, Max, et al.. "Bigger, Better, Faster: Human-level Atari with human-level efficiency" *arXiv preprint arXiv:2305.19452*, 2023. https://arxiv.org/abs/2305.19452.

Sutton, Richard S., and Andrew G. Barto. *Reinforcement Learning: An Introduction*. 2nd ed., MIT Press, 2018. http://incompleteideas.net/book/the-book-2nd.html.

Levine, Sergey, Aviral Kumar, George Tucker, and Justin Fu. "Offline Reinforcement Learning: Tutorial, Review, and Perspectives on Open Problems." *arXiv*, 2020, arXiv:2005.01643. https://arxiv.org/abs/2005.01643.

Kumar, Aviral, Aurick Zhou, George Tucker, and Sergey Levine. "Conservative Q-Learning for Offline Reinforcement Learning." *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 1179-1191. https://arxiv.org/abs/2006.04779.