How Aggregate Choices Change a Game Benchmark Comparison

Mark Brooks (AE), James Turner (GE), Raymond Spencer (CL)

Abstract

The main BBF file exceeds its comparison under several equally game-weighted location summaries, but the reported magnitude depends on how task tails are handled. Its mean of game means is 1.484, its median is 0.670, and removing thirteen games from each ordered tail leaves a mean of 0.748. Corresponding comparison values are 0.454, 0.189, and 0.244. These are different estimands, not interchangeable confirmations. The analysis uses the released scores to clarify how an interactive-rendering report can show aggregation sensitivity without selecting whichever summary produces the most appealing claim.

1. Aggregation is part of the scientific question

A benchmark score combines two decisions: which observations count and how their outcomes are summarized. Once a study has chosen to give every game equal influence, it still needs to decide whether its primary statement concerns the mean, median, or another feature of the task distribution. A trimmed mean changes that target by discarding observations from both tails. This can reduce sensitivity to extremes while also removing genuine difficult or unusually successful tasks from the statistic.

For interactive rendering, Make a game motivates evaluating a collection of controlled environments rather than treating every generated image as the complete task (Meng et al., 2026a). The present paper does not observe that system's outputs. It studies a public score archive in which alternative aggregations can actually be calculated. Its relevance is a reporting principle: a result should identify the task population and summarization rule before its magnitude is interpreted as evidence of broad capability.

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. Four declared summaries of the same game means

For each of the two official BBF CSVs, the analysis first computes a mean released normalized score within every game. Both configurations cover 55 games, so the subsequent comparison uses the same task labels with equal task weight. The four summaries are the ordinary mean, the median, the mean after removing five observations from each ordered tail, and the mean after removing thirteen from each tail. The last two retain 45 and 29 game means respectively. These exact counts define the operations more clearly than approximate percentage labels alone.

The main configuration has a mean of 1.4843, a median of 0.6701, a five-per-tail trimmed mean of 0.9934, and a thirteen-per-tail trimmed mean of 0.7476. The file without resets gives 0.4538, 0.1892, 0.3162, and 0.2440 under the corresponding operations. Thus the direction of this descriptive comparison survives all four choices. Its magnitude and practical description do not remain identical, because the summaries deliberately emphasize different parts of the distribution.

Replication has several levels: a second measurement, a second trajectory, and a second independently acquired setting are different additions to evidence. Statistical calculations should retain those distinctions. Heteroskedasticity-robust regression inference concerns unequal conditional variances (White, 1980); cluster-robust inference concerns dependencies shared within groups (Cameron and Miller, 2015). The group definition should be justified through the data collection and evaluation process, and not chosen because it produces a preferred significance result. Unequal group sizes also influence what a pooled average represents. Reporting both the group count and observation count permits a reader to see whether the calculation describes common cases, heavily recorded cases, or an explicitly balanced group population. No variance estimator alone resolves that choice of estimand.

A compact latent state is a hypothesis about which information an environment requires. World Models demonstrates one route to learning such a representation (Ha and Schmidhuber, 2018), and Dreamer connects learned latent dynamics to behavior (Hafner et al., 2019). IRIS and DIAMOND explore further world-model choices using transformers and diffusion respectively (Micheli et al., 2022) (Alonso et al., 2024). This literature makes information preservation a testable concern. If a renderer loses an object or confuses an action, the explanation could involve representation, history selection, conditioning, or downstream use. A meaningful comparison should separate those possibilities with matched inputs and explicit outcomes. The existence of a sophisticated latent representation does not establish that every relevant state variable survives compression. Nor does the ability to produce a convincing next image establish correctness over a long interaction. Those claims require observations at the corresponding temporal scale and a record of what the model was allowed to know.

The protected assessment population should be defined before investigators inspect detailed failures. Once a collection of cases guides repeated revisions, it has become development information even if its filenames still carry an evaluation label. Cross-validation for dependent data treats the grouping of observations as a substantive design decision (Roberts et al., 2017). Research on model selection likewise shows why tuning and final error estimation should not reuse the same information (Varma and Simon, 2006). A practical record therefore distinguishes fitting data, diagnostic examples, validation data, and the final external assessment. Where only a fixed public table is available, a descriptive analysis remains legitimate, but it should not be renamed a held-out prediction experiment without generating and assessing predictions under a protected split.

3. What robustness across summaries establishes

Agreement in direction is useful evidence that the observed ordering is not produced solely by one preferred location statistic. It is not four independent replications. Every calculation uses the same released experiments and mostly the same tasks. Their uncertainties are related, and their assumptions about the target population differ. An investigator should not combine them as though they were separate studies or use the most favorable one to replace a prespecified primary endpoint.

The extremes are scientifically informative as well. The main-file game means range from -0.8421 to 12.3364; the comparison range is -0.9998 to 6.0740. A trimmed summary reduces the influence of these cases but does not show whether they represent unusual reward scales, particular mechanics, or a different distribution of run outcomes. A task-level table remains necessary for inspecting those possibilities.

Trimming should therefore be presented as a sensitivity view rather than a procedure that cleans away inconvenient evidence.

In a direct rendering study, the treatment of extremes should follow the application. A rare but unacceptable control violation may deserve an explicit failure endpoint rather than removal from a trimmed visual-quality mean. An unusually strong output may be valuable for a best-of-many workflow but irrelevant to a single-shot experience. The scoring design should distinguish those use cases. A balanced benchmark average can describe a chosen game list while still failing to represent either workflow unless the observation unit and selection procedure are recorded.

The order of aggregation also matters. Averaging repeated outputs within each environment and then trimming environments discards different evidence from trimming outputs within every environment. The first changes the represented task population. The second changes the summarized outcome distribution within each task. Both operations can produce a quantity called a trimmed mean. A reproducible report should therefore state where the operation occurs in the hierarchy and retain intermediate results so the distinction can be checked. Here trimming is applied to the 55 game means, not to a pooled collection of all runs.

Control and fidelity should be assessed as related but distinct properties. Neural game generation in GameNGen (Valevski et al., 2024) and action-controlled environment generation in Genie (Bruce et al., 2024) bring both into view. Structural similarity (Wang et al., 2004) or a learned image distance (Zhang et al., 2018) may help quantify a reference comparison; a video-distribution metric (Unterthiner et al., 2018) can summarize another aspect of temporal appearance. These measurements do not identify whether the interface interpreted a command correctly. The evaluation record should link an output to the exact control history and to any available state observation. It should also distinguish permitted variation from a broken requirement. A stochastic rendering system need not reproduce every pixel to be useful, but it does need to preserve the aspects of interaction that its claim depends on. Measuring those requirements explicitly makes the relationship between visual quality and behavior open to examination.

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.

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. Interpreting playability evidence

An action-dependent observation has meaning only within a specified interaction process. Reinforcement learning provides a formal distinction among states, actions, rewards, and policies (Sutton and Barto, 2018). Offline learning further emphasizes that a fixed collection of histories has limited behavioral support (Levine et al., 2020). Conservative value learning addresses a particular problem arising when actions are evaluated beyond that support (Kumar et al., 2020). These sources do not validate a rendering architecture, but they clarify why familiar trajectories can be an incomplete test. A renderer may look reliable when replaying common histories while responding poorly to interventions that the archive rarely contains. A future evaluation should state which commands and transitions are supported by its data, which are deliberately held out, and which remain untested. Reweighting observed histories can alter their importance but cannot create an observation of a missing control consequence. That boundary should stay visible in the interpretation of any broad capability claim.

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.

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 comparison has the same direction under four declared task-level location summaries, while the size of the apparent advantage changes substantially. The appropriate conclusion is a bounded aggregation sensitivity, not universal superiority or repeated independent confirmation. Interactive-rendering evaluations should nominate a primary estimand, preserve tail cases, and show how alternative summaries answer different questions. A robust narrative is one that remains precise about what changes when the summarization rule changes.

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.