Game-Weighted Evidence for Interactive Rendering: A Reanalysis of Released Atari Scores

Oscar Nelson (NI), Mason Peterson (AZ), Austin Barnes (DO)

Abstract

A benchmark average can change simply because some games contribute more runs. In two released Bigger, Better, Faster score files, equal weighting of 55 games lowers the main configuration's mean normalized score from 1.832 to 1.484. A game-balanced comparison on a compressed score scale still favors that configuration, with an estimated difference of 0.466 and a game-bootstrap interval of 0.322 to 0.627. These are secondary calculations on published agent results. They motivate a practical distinction for interactive rendering research: report visual quality, control performance, and the distribution of evaluated environments separately.

1. An interactive world needs an explicit evaluation population

Interactive rendering changes what an evaluator must observe. A convincing frame is one event in a sequence conditioned on actions, previous images, and the environment's changing state. A renderer can maintain appearance while responding incorrectly to a control input, or reproduce short transitions while losing objects over a longer interaction. An overall score becomes meaningful only after specifying which games, histories, actions, and durations it represents. The question addressed here is narrower: how much does unequal representation of games change a summary of publicly released agent scores, and what does that imply for the design of rendering evaluations?

Make a game proposes interactive game rendering as its central research problem (Meng et al., 2026a). The official conference record establishes the work's identity, while the author's public description identifies unified tokens and lightweight plugins as elements of its approach. The complete publisher text was inaccessible during this analysis. Consequently, no architectural detail, latency measurement, training-data claim, or comparative performance number is reconstructed. The useful connection is an evaluation question: if an interface combines heterogeneous controls with generated imagery, evidence should establish both the fidelity of the display and the consequences of an action. The score analysis below supplies an example of how even a control endpoint depends on the population being averaged.

The Arcade Learning Environment established a broad platform for studying general agents across distinct games (Bellemare et al., 2012). OpenAI Gym standardized interaction interfaces (Brockman et al., 2016), and Procgen made procedural variation an explicit generalization challenge (Cobbe et al., 2019). Together they suggest that the environment is more than a row label. Different task populations change the distribution of mechanics, states, and opportunities to fail. None of these observations implies that an Atari return measures the quality of a newly generated game; they identify a useful precedent for specifying the conditions under which an interactive system is tested.

2. Data and comparison actually performed

The analysis uses two CSV files distributed in the official Bigger, Better, Faster repository: RR8_BBF and RR8_BBF-Resets. The associated work studies sample-efficient value-based learning and publishes score data alongside its implementation (Schwarzer et al., 2023). Here the files are treated as fixed released observations. No agent was trained, no emulator was run, and no rendering model was evaluated. The file names identify configurations, but they do not by themselves certify an independently randomized experiment or a common evaluation budget for every archived row. Interpretation therefore remains a comparison within the released files.

The main file contains 2,082 records and the configuration without resets contains 830. Both cover the same 55 game labels, and the combination of game, seed, and configuration is unique in each file. There are no missing or nonfinite normalized scores. Run counts range from 16 to 60 per game in the main file and from 15 to 20 in the other. Thus a record-weighted average and an equally weighted average of game means answer different questions before any statistical model is fitted.

The released GameScoreNormalized field is used as supplied. The analysis does not reconstruct normalization constants or infer comparability with other archives. A second outcome applies the inverse hyperbolic sine transformation to each normalized score. This transformation accommodates zero and negative observations while reducing the leverage of extremely large positive scores. Its coefficient is a difference in transformed units, not a percentage gain, a difference in visual fidelity, or a causal treatment effect.

The distinction between a population of tasks and repeated runs is central to reliable reinforcement-learning evaluation (Agarwal et al., 2021). It is also compatible with the separation of states, actions, and returns in reinforcement learning (Sutton and Barto, 2018). The unit being averaged should follow the scientific question. If the question concerns a uniformly selected game from this archive, additional runs should improve information about that game's distribution without automatically increasing its contribution to the target population.

3. Weighting, uncertainty, and the estimand

Two descriptive means are calculated for each configuration. The pooled mean gives every recorded run equal weight. The balanced mean first averages runs within a game, then averages the 55 resulting game means. Neither is universally correct. The pooled result describes this archive's run allocation; the balanced result describes its equally weighted game list. Their difference is a direct sensitivity calculation rather than a correction justified by a presumed random sampling mechanism.

For the main comparison, weighted least squares regresses the transformed score on a configuration indicator and game fixed effects. Each row receives the reciprocal of its configuration-by-game run count. Each configuration therefore has equal total weight within each game. The common coefficient equals the mean of the game-specific transformed contrasts in this balanced two-configuration design. Repeated runs share their game, so uncertainty is clustered by the 55 game labels, with a finite-cluster adjustment and a t reference distribution (Cameron and Miller, 2015). Heteroskedasticity-consistent covariance alone would not address this grouped structure (White, 1980).

A second interval resamples the 55 game-specific contrasts with replacement 5,000 times, using seed 461. This is a percentile bootstrap over games. It represents sensitivity to the empirical mixture of game contrasts and is not a new set of training runs. The result depends on treating the observed tasks as a resampling population. It cannot establish uncertainty over all possible games, unseen mechanics, or newly generated worlds. Keeping that interpretation visible avoids making an attractive confidence interval answer a broader question than the data support.

The model is descriptive. A causal diagram would need to represent training resources, implementation changes, task selection, and evaluation procedures before interpreting a configuration coefficient as the isolated effect of resets (Greenland et al., 1999). The files do not supply a verified intervention assignment covering those factors. Clustered precision concerns an estimated archive contrast; it does not repair identification.

4. What changes when games receive equal weight

Configuration	Recorded runs	Pooled mean	Equal-game mean
Main BBF file	2,082	1.832	1.484
File without resets	830	0.452	0.454

The main configuration's balanced mean is about nineteen percent below its pooled mean. The alternative configuration changes little under the same operation. This asymmetry indicates that unequal representation is consequential in the first file. It does not identify why more runs appear for particular games. A release may combine experiments with different coverage; the present calculation documents the consequence without inferring a hidden selection policy.

Dispersion remains substantial after balancing. The median of the main configuration's game means is 0.670, compared with its equal-game mean of 1.484. For the file without resets, the corresponding median and mean are 0.189 and 0.454. Means therefore summarize distributions with influential upper tails. The median is informative about location, but it discards the magnitude of many gains and losses. Reporting both is more useful than presenting either as a complete description of robustness.

The balanced transformed contrast is 0.466, with a game-clustered standard error of 0.083 and a 95 percent interval from 0.300 to 0.633. The game-bootstrap interval is 0.322 to 0.627. Of the 55 game-specific transformed contrasts, 49 are positive, five negative, and one zero. Agreement between these summaries supports a broad within-archive advantage, while the negative contrasts show that the average does not imply improvement on every task. Full game-level results are retained with the analysis, so a reader can examine the distribution rather than depend on a selected example.

The bootstrap and regression intervals are not independent replications: both use the same released records. Their proximity shows that the conclusion is not solely an artifact of one interval calculation. No hyperparameters or game subsets were tuned to produce the reported coefficient. If alternative transformations or exclusion rules are compared in future work, those choices should be documented before selecting a preferred result.

5. Transferring the lesson to generated interaction

World-model research provides several ways to connect a representation to behavior. World Models explores compressed representations of environment experience (Ha and Schmidhuber, 2018), while Dreamer learns behaviors through latent imagination (Hafner et al., 2019). IRIS investigates transformer-based world models (Micheli et al., 2022), and DIAMOND studies diffusion-based modeling in which visual detail matters to Atari interaction (Alonso et al., 2024). These approaches make the connection between predicted observations and downstream decisions scientifically important. The present archive contains agent scores, however, and cannot rank these world-model architectures.

Generative interaction also has distinct objectives. GameNGen studies a neural model supporting real-time interaction (Valevski et al., 2024), whereas Genie learns action-controllable environments from unlabeled videos (Bruce et al., 2024). Their evaluation questions help situate Make a game: response to input, continuity over time, and the range of worlds supported should each be observable. Shared terminology does not make their datasets or scores interchangeable. A proposed renderer should state whether its controls correspond to known game actions, a learned action vocabulary, text commands, or an interface assembled by plugins.

Visual endpoints should accompany a control endpoint. Structural similarity compares image structure under defined reference conditions (Wang et al., 2004); learned perceptual similarity represents a different relationship between image pairs (Zhang et al., 2018). Video-distribution evaluation introduces temporal information and its own dependence on representations and sample size (Unterthiner et al., 2018). None of these metrics directly answers whether a delayed action had the intended consequence. An evaluation record should preserve the action history and the reference state when those quantities are available, rather than combining every endpoint into one unexplained total.

The distribution of available histories matters as much as the choice of a visual metric. Offline reinforcement learning makes the limitation of learning from a fixed behavioral distribution explicit (Levine et al., 2020), and conservative value learning addresses a particular aspect of out-of-distribution action evaluation (Kumar et al., 2020). Those methods are not applied here. Their relevance is that a renderer evaluated only on familiar action sequences can appear reliable without having demonstrated useful behavior under alternative interventions. Data coverage and response quality should be reported together.

6. A reproducible next evaluation

A practical study would nominate a finite set of games, allocate the same number of evaluation histories to each, and retain both game-level and history-level results. A second design could target actual user traffic; that would require a documented traffic distribution and would justify different weights. The choice should precede aggregation. When histories share a scene, player, or recording, train-test boundaries should follow those dependencies (Roberts et al., 2017). If model selection is involved, its validation boundary must also remain separate from the final assessment (Varma and Simon, 2006).

Reliability requires more than ranking. Calibration asks whether stated uncertainty corresponds to observed frequencies (Guo et al., 2017), while predictive uncertainty under shift may behave differently from uncertainty within the original data distribution (Ovadia et al., 2019). A renderer permitted to decline unsupported commands should report its coverage alongside conditional performance, following the distinction formalized in selective prediction (Geifman and El-Yaniv, 2017). These are proposed evaluation additions; the released score files contain neither confidence forecasts nor abstention decisions.

Metric agreement should be measured rather than assumed. Two scores can correlate while disagreeing systematically for particular games; the method-comparison distinction is relevant when replacing a human or task endpoint with an automated proxy (Bland and Altman, 1986). When many game, duration, and metric comparisons are reported, multiplicity should be addressed instead of highlighting whichever comparison looks strongest (Benjamini and Hochberg, 1995). Stable identifiers for inputs, configurations, and results make the evaluation reusable (Wilkinson et al., 2016), and explicit tracking of dependencies reduces the risk that an evaluation pipeline quietly changes its meaning (Sculley et al., 2015).

7. Conclusion

The same released BBF runs produce a pooled mean of 1.832 and an equal-game mean of 1.484. Balancing games does not remove the observed advantage over the comparison file, but it changes its magnitude and clarifies the population represented. For interactive rendering, the corresponding requirement is to state which environments and histories receive weight, preserve their results, and keep visual evidence connected to action consequences. The analysis establishes an aggregation sensitivity in public agent scores; a rendering system's fidelity, latency, and controllability still require direct observation.

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.