Where a Configuration Loses: Heterogeneity across Released Atari Tasks

Alexander Mason (KW), Robert Bailey (GN), Michael Rogers (CZ)

Abstract

An average improvement leaves open the question of where a system becomes worse. In the released BBF comparison examined here, 49 of 55 game-level contrasts favor the main configuration, five favor the file without resets, and one is zero. The median transformed contrast is 0.205, less than half its mean of 0.466. Losses are relatively small on the chosen scale, while the strongest positive contrast reaches 2.878. These descriptive findings motivate reporting the distribution of task outcomes when evaluating interactive rendering. They do not establish an improvement in any renderer or isolate the causal effect of a component.

1. A population average does not identify the supported task

A user interacts with a particular environment, not with the mean of a benchmark. An aggregate can describe a broadly successful configuration while concealing a small set of environments in which it performs less well. That discrepancy need not invalidate the average. It requires a second statement about the distribution of outcomes. A research claim becomes clearer when it distinguishes how large an average advantage is from how consistently the advantage occurs across the represented tasks.

Make a game places controllable rendering at the boundary between generated imagery and interactive environments (Meng et al., 2026a). The author's public account mentions unified tokens and lightweight plugins. Those elements invite an evaluation organized around particular tasks and control contexts: a plugin may work in one setting and fail in another. The publisher text was unavailable for this study, so no unsupported details about its implementation or benchmark performance are supplied. The present contribution is an analysis of heterogeneity in a separate public agent-score archive that informs how an interactive evaluation could be reported.

The Arcade Learning Environment makes varied games available within a common evaluation platform (Bellemare et al., 2012). Its common interface does not make the tasks equivalent. Differences in reward opportunities, mechanics, and required histories remain part of the scientific population. Reinforcement learning formalizes behavior through states, actions, and returns (Sutton and Barto, 2018); the task-specific interpretation of those quantities matters before an average can be translated into a claim about a particular interactive world.

2. A released archive rather than a new agent experiment

Two score CSVs from the official Bigger, Better, Faster repository supply the observations (Schwarzer et al., 2023). The main file contributes 2,082 rows and the file labeled without resets contributes 830. Both include 55 game labels. Each configuration-game-seed combination is unique. The records are treated as the released results of their original experiments, with no training or evaluation rerun here. File labels do not establish a newly randomized component experiment or a complete account of all protocol choices.

The outcome is the supplied normalized score. Within each configuration and game, the analysis calculates the mean of the inverse hyperbolic sine of the score. This transformation retains negative and zero observations and compresses very large positive values. The difference between the two transformed means is the game-level contrast. A positive contrast favors the main file; a negative contrast favors the comparison file. Its units are transformed score units, not percentage points of rendering fidelity.

This construction gives each task one contrast irrespective of its run count. It answers a distributional question about the represented game list. A run-weighted analysis would emphasize games with more archived experiments and would be a legitimate answer to a different question. Reliable benchmark evaluation requires making that target explicit instead of choosing an average only after observing which one looks most favorable (Agarwal et al., 2021).

All 55 contrasts are retained. There is no threshold-based deletion of inconvenient games, no selection of examples through statistical significance, and no new environment data. The analysis does not estimate how frequently a task would be encountered by a player. Such an interpretation would require an external distribution over use conditions. Equal task weighting is a transparent benchmark convention, not a description of user traffic.

3. The direction and size of the task contrasts

The main configuration has a positive transformed contrast on 49 games, a negative contrast on five, and an exactly zero contrast on one. Performing the directional check on untransformed game means gives the same counts. The sign pattern is therefore not produced solely by compressing the upper tail. It remains a description of these files, because the underlying configurations and experiments have not been reconstructed as a controlled intervention.

Summary of transformed game contrasts	Value	Population	Meaning
Minimum	-0.0904	55 games	Largest observed loss
Lower quartile	0.0440	55 games	A modest positive contrast
Median	0.2051	55 games	Middle task contrast
Upper quartile	0.6966	55 games	Larger positive contrasts
Maximum	2.8778	55 games	Largest observed gain

The mean contrast is 0.4662. Its distance above the median indicates that a statement about the average magnitude is influenced by larger positive outcomes. That does not imply that most games fail to improve: the sign count shows the opposite. Magnitude and breadth are separate summaries, and both are needed to avoid an exaggerated or unduly pessimistic interpretation.

The five negative contrasts occur in Bank Heist, Bowling, Private Eye, Solaris, and Venture. Their transformed values range from approximately -0.0904 to -0.0058. Identifying them permits inspection, but it does not establish a common mechanism. The files contain scores, not classified failure trajectories. Inferring that the losses arise from memory, exploration, or rendering instability would require observations not available in this reanalysis.

Games are not exchangeable merely because a common interface launches them. OpenAI Gym illustrates the utility of standardized interaction interfaces (Brockman et al., 2016), while Procgen explicitly investigates procedural variation within environment families (Cobbe et al., 2019). Together they suggest reporting where a test set gets its diversity. A count of 55 tasks is a useful description, but it does not guarantee coverage of every mechanic relevant to a proposed generative interface.

A useful negative-task dossier would preserve both the common and unusual features of each recorded episode. Suppose a method loses on a game containing sparse rewards. That association can motivate a hypothesis, but sparse rewards may accompany long exploration paths, rare terminal events, or a different

balance of observations. A subsequent experiment should vary the proposed difficulty while controlling the others where feasible. It should also include games outside the original negative set. Otherwise, the investigator risks explaining a small collection retrospectively with a feature that was selected because it happens to fit those examples. The task table is therefore a starting point for controlled inquiry. It does not supply an explanation simply by allowing an evocative label to be attached to a loss.

4. Why the observed losses are not a mechanism test

A configuration difference can combine changes in training, representation, optimization, evaluation conditions, and data allocation. A causal diagram would need to describe those relationships before the game contrasts could be interpreted as the isolated effect of one component (Greenland et al., 1999). Naming the removed component does not supply the missing assignment mechanism. The present distribution therefore supports a descriptive statement about released outcomes and an agenda for more controlled comparisons.

Uncertainty also has a grouping structure. Runs associated with the same game share its reward scale and mechanics. A model that treats all rows as independent cannot automatically supply an interval for the average task contrast. Cluster-robust inference provides tools for grouped observations (Cameron and Miller, 2015), whereas heteroskedasticity-consistent covariance alone addresses a narrower variance issue (White, 1980). Neither technique turns a selected benchmark into a probability sample of all possible interactive tasks.

There is also a distinction between testing whether a mean contrast differs from zero and estimating the fraction of tasks improved. A positive mean can coexist with many losses if a few gains are large. Conversely, a broadly positive sign pattern can coexist with a small mean if improvements are tiny. The current data show broadly positive signs and a right-skewed magnitude distribution. A future report should preserve both rather than translate one into the other.

Multiple task comparisons can tempt an investigator to narrate only their most interesting outcomes. False-discovery procedures formalize one response when a family of inferential tests is actually conducted (Benjamini and Hochberg, 1995). Here the five negative tasks are listed descriptively, with no task-specific significance claim. The absence of p-values should not be mistaken for an invitation to label every observed difference a confirmed effect.

5. Applying heterogeneity analysis to generated environments

World Models learns compressed representations of environment experience (Ha and Schmidhuber, 2018), and Dreamer connects learned latent dynamics with behavior learning (Hafner et al., 2019). IRIS investigates transformer-based world modeling (Micheli et al., 2022), while DIAMOND studies diffusion representations that retain behaviorally relevant visual detail (Alonso et al., 2024). These approaches indicate several routes by which a task-specific weakness can arise. The present agent results cannot discriminate among those routes because no corresponding model outputs or state trajectories are analyzed.

For interactive rendering, a natural analogue is to define a contrast for each environment under a matched action-history protocol. GameNGen studies a neural model capable of real-time game interaction (Valevski et al., 2024), and Genie develops action-controllable environments from unlabeled video (Bruce et al., 2024). Comparisons involving such systems should specify whether the same controls, reference histories, and evaluation horizons are available. Otherwise, a task contrast may combine an input advantage with a difference in model quality.

Visual outcomes should also remain separate from returns. Structural similarity assesses defined image-pair structure (Wang et al., 2004), learned perceptual similarity measures another aspect of image correspondence (Zhang et al., 2018), and video-distribution metrics bring temporal features into generative evaluation (Unterthiner et al., 2018). A task that improves on a visual metric can still respond incorrectly to an action. Conversely, task return can improve without a corresponding increase in image quality. Reporting a multivariate task profile is more informative than assuming these endpoints are interchangeable.

Offline learning emphasizes that behavior is constrained by the support of the available experience (Levine et al., 2020). Conservative value learning addresses one particular out-of-distribution problem (Kumar et al., 2020). The parallel for a renderer is that familiar histories may conceal task weaknesses that become visible under unusual interventions. A follow-up test should include controlled action changes and distinguish an unsupported history from a history on which the system gives a confidently wrong response.

6. Turning losses into a useful research design

The first step is a failure record containing the environment, initial state, action sequence, duration, and all primary outcomes. It should permit a second observer to reproduce what was being compared. Stable identifiers and reusable provenance make that record easier to interpret (Wilkinson et al., 2016). Explicit model and scoring dependencies are necessary because a change in preprocessing or metric version can alter which tasks appear weak (Sculley et al., 2015).

The next step is to protect the evaluation boundary. Tasks or histories used to diagnose a failure should not silently become part of a model's final held-out assessment. Structured validation preserves dependencies among related observations (Roberts et al., 2017), and model-selection research shows why tuning on the assessment set biases error estimates (Varma and Simon, 2006). A convincing repair would be tested on new histories exhibiting the hypothesized difficulty, rather than only on the examples that inspired the change.

Reliability can also involve a decision to defer. Calibration concerns whether stated probabilities correspond to observed frequencies (Guo et al., 2017), while uncertainty under distribution shift may behave differently from familiar-condition uncertainty (Ovadia et al., 2019). A renderer allowed to reject an unsupported control should disclose how often that happens and how accepted interactions perform (Geifman and El-Yaniv, 2017). Such reporting prevents the appearance of universal task coverage when the system actually succeeds by restricting its scope.

Finally, an automated task score should be compared with the intended human or behavioral outcome. The distinction between correlation and agreement is relevant when one measurement replaces another (Bland and Altman, 1986). In a rendering study, users may tolerate a small visual deviation that a metric penalizes, yet reject a control error that leaves the image almost unchanged. The interpretation of task-level losses depends on observing those consequences rather than relying only on a convenient proxy.

7. Conclusion

The released comparison favors the main configuration on most represented games, but five observed losses and a skewed contrast distribution remain visible when task outcomes are retained. The median task difference is substantially smaller than the mean. An interactive rendering evaluation should adopt the same distinction between breadth, magnitude, and mechanism. A distribution of outcomes can identify where further evidence is needed; explaining those outcomes requires controlled histories and measurements beyond a table of returns.

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.