Replaying an Interaction when the Renderer Is Stochastic

Evan Morris (SG), Frank Graham (DO), Julian Webb (MY)

Abstract

Replay can mean reproducing bytes, preserving a task consequence, or recovering a comparable user experience. These are different requirements for a stochastic generative interface. This conceptual study proposes a replay record linking the initial history, controls, model version, randomness configuration, and evaluator. It also distinguishes acceptable visual variation from changes that invalidate the interaction. Make a game motivates the rendering setting, but no new replay experiment is performed. The design allows a future study to state exactly what it has reproduced instead of treating a shared seed label as a complete reproducibility guarantee.

1. Repetition has several possible targets

A user may want to repeat an interaction to inspect a failure, compare a revised model, or share an experience. Exact pixel reproduction would serve some of those purposes, but it is not always necessary. A stochastic renderer could produce different textures while preserving object states and action consequences. Conversely, two visually similar clips could differ in a crucial event. The replay specification should identify which properties are required to remain stable and which are permitted to vary.

Make a game frames generation as part of an interactive interface (Meng et al., 2026a). Its public description motivates the question of replay, but does not establish a complete state-saving or randomness protocol. This article therefore develops a research design rather than reporting a feature of the implementation. The proposed distinction is between computational repeatability and task-level replay. Either can be useful, provided the evidence supports the particular requirement a report claims to satisfy.

A benchmark interface supplies a common language for launching tasks, but it does not eliminate differences among the tasks. The Arcade Learning Environment was designed to evaluate general agents across games (Bellemare et al., 2012). Gym offers a common interaction interface (Brockman et al., 2016), while Procgen makes procedural variation an explicit part of the assessment population (Cobbe et al., 2019). These contributions help distinguish software compatibility from evidence of transfer. The BBF release supplies actual repeated-run scores that can be reanalyzed (Schwarzer et al., 2023), whereas the statistical-precipice analysis explains why uncertainty and aggregation choices deserve attention (Agarwal et al., 2021). Together, these sources motivate preserving task-level results. They do not imply that an Atari return is a visual-quality score or that a renderer supports the mechanics of every game available through the same interface. A proposed evaluation should identify its games, initial conditions, and action histories, then state which population the aggregate represents. Additional runs can improve knowledge of a task's outcome distribution without increasing the number of different tasks observed.

2. Recording the conditions that a replay depends on

A replay package should identify the initial observation or state snapshot, the prior history available to the model, and the ordered control sequence. It should also preserve the model checkpoint, plugin versions, preprocessing, and sampling configuration. A random seed belongs in that record, but its scope must be explained. It may control only one sampling stage while other software, scheduling, or input processes contribute additional variation. Reusing the same integer does not by itself demonstrate that the experiment was repeated under the same effective conditions.

The output record should distinguish generation from scoring. A stochastic evaluator can make a fixed generated clip receive different judgments across runs. A changed score then need not reflect a changed renderer. Keeping the evaluator version and invocation separate permits this possibility to be checked. Where exact execution is impractical, the release can still define a bounded semantic replay target and retain enough observations to assess it. That is a scientific design choice rather than a reason to imply exact reproducibility that has not been demonstrated.

The proposed test would include repeated invocations of fixed histories, with deliberate variation of nominated randomness sources. It would report whether required states and action consequences persist, how visual quality varies, and how often a valid output or measurement is unavailable. A single favorable replay would not establish a stable distribution of outcomes. Nor should an investigator discard visually different but task-valid continuations merely because they do not match a selected reference frame.

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.

World modeling is not defined only by the visual appearance of predicted frames. It concerns the relationship between experience, representation, and subsequent behavior. World Models (Ha and Schmidhuber, 2018) and Dreamer (Hafner et al., 2019) illustrate compressed and latent-imagination approaches, while IRIS (Micheli et al., 2022) and DIAMOND (Alonso et al., 2024) investigate different sequence-modeling mechanisms. These examples caution against evaluating every system with a single generic image metric. A control task can depend on a small visual event, a hidden counter, or the persistence of an object outside the current view. An experiment should specify which of those quantities is observable and which is only hypothesized. If the underlying state cannot be recorded, a more limited behavioral contrast may still be informative. The important step is to connect the architecture's proposed advantage to a consequence that could be supported or contradicted by the actual evaluation.

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.

3. Comparing a repaired renderer with its earlier version

A replay record is particularly valuable when a model is updated. The original failure history can become a diagnostic case, but success on that case after repair is not independent evidence of general improvement. A stronger study would retain a protected collection of related but unexamined histories. It would compare the old and new systems under a fixed task definition and declare whether the pairing shares initial evidence, actions, or another source of variation. The scientific pairing should follow those shared objects.

Some revisions may intentionally change the task. A plugin can adopt a new precedence rule, or an interface can narrow the supported command set. In that case, an old replay may no longer be eligible under the revised contract. The report should identify the change rather than label every incompatible output a regression. At the same time, narrowing eligibility should not be presented as a pure gain in quality. Both versions' support boundaries and accepted-case performance are needed for a meaningful comparison.

Reproducibility also has an access dimension. A public result may depend on a checkpoint or reference asset that cannot be redistributed. A release can document that dependency, provide hashes, and state which parts can be rerun independently. It should not imply that a video demonstration supplies the missing execution environment. The proposed record makes such limits explicit while preserving the observations that remain useful for later inspection and controlled comparison.

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.

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.

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. Reliable evidence about the interaction process

An interactive benchmark should explain how its histories were obtained. In reinforcement learning, the behavior policy and the environment jointly determine which observations appear (Sutton and Barto, 2018). Offline-learning research highlights why fixed histories leave some actions and states weakly supported (Levine et al., 2020); conservative value learning provides a concrete example of handling a related estimation difficulty (Kumar et al., 2020). For a renderer, this means that a favorable replay score may be specific to the histories an archive happens to contain. A stronger test would deliberately vary a control while preserving the relevant initial evidence. It would then measure whether the expected consequence occurred and whether visual quality remained acceptable. Such a study requires a valid reference or task rule. Without that reference, a secondary analysis can describe available coverage and score behavior but cannot establish unrestricted correctness of the generated world's dynamics.

Coverage should be reported whenever the system or its evaluator can avoid difficult cases. An abstention policy changes the population on which an accuracy or fidelity score is calculated. Selective classification formalizes the joint reporting of coverage and conditional performance (Geifman and El-Yaniv, 2017). Distribution-shift research adds that estimated uncertainty can behave differently outside the development population (Ovadia et al., 2019). A benchmark should therefore retain which observations were eligible, accepted, rejected, or unavailable. A result on a carefully limited subset can be useful, provided the subset is visible. The challenge is to prevent a narrow success claim from becoming a broad capability claim through omission of the denominator or the circumstances in which the system declines to operate.

A scientific release should make its most consequential choices recoverable. Persistent identifiers and explicit metadata support reuse of observations and results (Wilkinson et al., 2016), while attention to dependencies helps prevent an evaluation from changing unnoticed as software evolves (Sculley et al., 2015). Relevant choices include the source version, preprocessing, accepted input set, metric implementation, and weighting convention. The record should also separate an observed output from a proposed future test. Saving that distinction prevents a plan from later being mistaken for a completed experiment. A small collection of well-linked records can be more useful than a large folder of unexplained scores. The important property is whether an independent reader can reconstruct what was measured and why its reported number has that meaning.

5. Conclusion

Replay should state whether it preserves exact computation, required consequences, or a declared class of acceptable experiences. A recoverable record links the initial history, controls, versions, randomness, and scoring decisions. Repeated observations then assess the relevant stability instead of assuming it from a seed. This article proposes that design without claiming a completed replay test. The resulting distinction makes stochastic variation compatible with honest reproducibility while keeping task-changing errors visible.

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.