Versioned Evidence for Interactive Rendering Systems

Austin Elliott (TH), Patrick Gardner (RS), Arthur Parker (CM)

Abstract

A model release can change the renderer, its plugins, the evaluator, and the supported task set at the same time. This conceptual study proposes a versioned evidence record that separates those changes and preserves their consequences for comparison. Make a game motivates an interactive interface with multiple moving components. No release or deployment experiment is conducted here. The argument is that progress should be assessed under a recoverable task definition: a higher score is difficult to interpret when the accepted inputs, reference histories, or measurement rule have also changed.

1. A new score can describe a new question

When an interactive system is updated, users often want a concise statement that it has improved. Yet several changes can contribute to the reported result. The generator may produce better outputs, a plugin may reinterpret a control, the benchmark may exclude unsupported cases, or the evaluator may use a new representation. A single score can combine all of these changes. Unless their identities are preserved, an apparent improvement may concern a different scientific task rather than better performance on the original one.

Make a game motivates this issue through its interactive rendering setting and public description of lightweight plugins (Meng et al., 2026a). That description does not establish a specific release process, so one is not inferred. The present article develops a conceptual record for evaluating an evolving interface. Its purpose is to separate a renderer change from changes in task support and measurement. A versioned record makes those distinctions reviewable without requiring every release to reproduce an identical implementation forever.

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. The scientific objects that need a version

A proposed evidence record would identify the task specification, initial histories, command definitions, model checkpoint, plugin configuration, and scoring implementation. These objects need separate identities because they can change independently. A benchmark name alone cannot establish that two releases use the same histories or validity rules. The record should also preserve the eligible input set and the reasons cases are accepted, rejected, or unavailable. This permits a comparison to distinguish broader support from better conditional performance.

The evaluator is especially important. Changing a feature extractor, threshold, or reference set can alter scores while the generated clips remain fixed. A useful release could rescore a saved set of outputs under both evaluator versions, separating measurement drift from generation changes. That would be an actual empirical comparison if performed; it is only a proposed design here. Saving per-case observations makes such a study possible without recreating every expensive output or relying on undocumented aggregate values.

Task revisions also require explicit interpretation. A plugin may adopt a new precedence rule to make an ambiguous combination well defined. Old reference outputs may then cease to be appropriate. The study can compare each version under its own contract, but it should not describe the difference as a pure quality gain under an unchanged task. Alternatively, a fixed common subset can support a narrower direct comparison, accompanied by a separate account of support expansion or contraction.

The choice of statistical unit should precede the choice of a standard error. A dataset can contain many rows but relatively few independently acquired histories. Heteroskedasticity-consistent covariance protects an otherwise specified regression against unequal residual variances (White, 1980). It does not automatically accommodate observations sharing an environment, subject, or recording. Cluster-robust inference makes that dependence part of the variance calculation (Cameron and Miller, 2015). The grouping must follow the observation process and the claim being made; an identifier that is convenient to store is not necessarily the relevant independent unit. For a descriptive analysis, preserving the groups also allows readers to recalculate a different population average. An uncertainty interval remains conditional on the observations and assumptions actually available.

Representation choices can change what an interactive system remembers and what it discards. The compressed environment representations in World Models (Ha and Schmidhuber, 2018), latent-imagination approach in Dreamer (Hafner et al., 2019), transformer world model in IRIS (Micheli et al., 2022), and diffusion formulation studied by DIAMOND (Alonso et al., 2024) offer different examples. Their differences matter because an apparently small visual feature can determine a later action or reward. A proposed rendering mechanism should be assessed through controlled observations of that consequence, rather than justified solely through the family of architecture it resembles. It should also distinguish short-horizon reconstruction from the persistence of behaviorally relevant information. No comparison among those architectures is performed by describing their designs. Their role here is to identify alternative explanations and appropriate baseline questions for a subsequent study. An informative experiment would match available history, training access, action information, and evaluation duration before attributing a difference to one representation.

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. Diagnosing progress across successive releases

A proposed report would present three views when the evidence permits: results on a frozen common assessment, results under the new supported task population, and an accounting of pipeline changes. The first supports a controlled comparison, the second describes the current operating scope, and the third explains why their denominators differ. None should be selected solely because it gives the most favorable headline. The comparison should retain failures and missing measurements under each declared rule.

Diagnostic cases create a further boundary. A failure discovered in an earlier release can guide repair, but success on that exact case is development evidence. An external assessment needs related cases not used to choose the repair. The record should therefore identify which histories have become diagnostic and when that status changed. A protected evaluation set can remain useful only if repeated consultation is acknowledged and managed. Changing filenames or release labels does not restore information that is no longer unfamiliar.

Versioning should remain proportionate. The goal is not a large inventory of irrelevant software details, but a recoverable account of decisions that can alter the task, outputs, eligibility, or scores. Where an asset cannot be shared, its identity and access limitation should be documented. Where a source cannot be rerun, saved outputs and verified hashes can still support some comparisons. The report should state which level of reproducibility is available rather than imply that a public demonstration supplies a complete executable study.

An interactive video system can fail while every individual frame remains plausible. GameNGen (Valevski et al., 2024) and Genie (Bruce et al., 2024) illustrate why action-conditioned generation deserves an evaluation distinct from unconstrained video production. The standard visual toolbox includes structural similarity (Wang et al., 2004), learned perceptual metrics (Zhang et al., 2018), and temporal distribution measures (Unterthiner et al., 2018). Their value depends on the reference information and feature representation being used. A high score cannot certify an unmeasured consequence such as correct collision behavior, object permanence, or response to a counterfactual command. The report should therefore retain separate outcomes and inspect their disagreements. If a visual metric is used as a proxy for playability, that substitution needs direct evidence. A plausible next step is a matched-history study in which control decisions and visual judgments are observed together, with durations and failure definitions fixed before examining the preferred model's outputs.

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 substitute measurement needs evidence about its relationship to the intended judgment. Two scores may correlate strongly while differing systematically in particular conditions. Method-comparison analysis distinguishes association from agreement (Bland and Altman, 1986). When a system additionally reports confidence, calibration asks whether those probabilities correspond to observed outcome frequencies (Guo et al., 2017). These are separate properties: a metric can rank examples usefully while having a

distorted numerical scale or unreliable confidence. An evaluation should therefore retain the original target observations when possible, describe the scale of each endpoint, and examine disagreements instead of averaging them away. A convenient automated score becomes more credible when its limitations are measured in the conditions that motivate its use.

4. Controls, histories, and the evaluation boundary

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.

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.

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

Progress in an interactive rendering system is easier to interpret when the task, generator, plugins, and evaluator have separate recoverable identities. The proposed evidence record supports fixed-task comparisons while making changes in scope visible. It also distinguishes diagnostic repair from independent assessment. No release outcome is claimed here. The contribution is a practical framework for ensuring that a higher score describes a declared improvement rather than an undocumented change in what the system was asked to do or how its outputs were judged.

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.