Counterfactual Controls for Testing Whether a Generated Game Responds

Peter Newman (CL), Connor Anderson (TM), Adam Bennett (MN)

Abstract

A video can change after a command without changing because of it in the intended way. This research design proposes matched-history comparisons between alternative controls, including a declared no-action condition where it is meaningful. Outcomes would assess required state consequences as well as visual quality and unsupported inputs. The proposal is motivated by Make a game's interactive rendering problem. No counterfactual generation experiment is reported. Its contribution is to specify the observations needed to separate plausible continuation, command recognition, and correct intervention in a stochastic visual environment.

1. A response needs an alternative

If an object moves after a user issues a command, the observed sequence alone may not establish that the command controlled the motion. The object might have moved under an ongoing process, a prior action, or a learned visual continuation. An interactive claim is stronger when the evaluation compares what happens under alternative inputs from sufficiently matched initial evidence. This does not require identical pixels. It requires a task rule that distinguishes the consequences of the alternatives.

Make a game addresses generated rendering under interaction (Meng et al., 2026a). A public description of unified tokens and control plugins suggests an interface through which alternative commands can be specified. It does not establish a causal evaluation protocol or an observed counterfactual effect. The present article develops that protocol as a proposal. The objective is to make the meaning of control testable while recognizing that a stochastic generator can support several acceptable visual realizations of the same required consequence.

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.

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.

2. Matching the history that defines the comparison

A matched-history test would retain the same initial observation, relevant prior actions, model checkpoint, and declared generation settings, then vary a nominated control. The comparison should state whether random sampling is shared, independently repeated, or otherwise managed. A common seed label alone would not guarantee that all stochastic events remain matched after the control changes. The scientific pairing should follow the information and processes actually held constant, rather than a convenient numeric identifier.

The task rule should identify which consequence is expected to differ. Turning left versus right might concern orientation or position; interacting with an object might concern a discrete state transition. A no-action condition can be informative only when the environment specification gives it a meaningful interpretation. It should not be assumed that absence of a command freezes every ongoing process. A moving environment can continue to change correctly while the user remains idle. A valid negative control therefore depends on the task's dynamics and interface semantics.

Repeated outputs would permit estimation of response distributions rather than a comparison between two selected clips. The study should retain every scheduled invocation, including failures to generate, unsupported commands, and outputs that cannot be measured. A visible difference between the two conditions is not enough: the difference must correspond to the required control consequence. A generator that changes color under a steering command is responsive in a superficial sense but has not demonstrated steering.

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.

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.

An evaluation protocol should state what remains unfamiliar at assessment time. Leaving out an observation while retaining its close relatives can test interpolation rather than the broader transfer

suggested by the study's wording. Validation methods for structured data emphasize matching the split to temporal, spatial, or hierarchical dependence (Roberts et al., 2017). Separately, parameter selection can bias an evaluation if the same observations guide both choices and reporting (Varma and Simon, 2006). These issues are not solved by calling a split a test set. The record should identify shared sources, the information accessible during development, and the stage at which each design choice was fixed. A result can be informative within a narrow boundary while providing no direct evidence about a more distant deployment population.

3. Separating command fidelity from visual plausibility

A useful outcome profile would contain at least an action-consequence endpoint, a visual-quality endpoint, and coverage. The first evaluates the task rule, the second evaluates appearance, and the third records how many intended cases yield valid observations. Their disagreements can expose important tradeoffs. A visually attractive sequence might ignore the intervention, while a correct transition might introduce artifacts. Combining the endpoints before inspecting those disagreements could hide which capability a model actually improves.

The design should also probe delayed effects. An action can have a correct immediate response but an incorrect later consequence. Histories should therefore include a declared observation horizon and, when relevant, subsequent actions that reveal whether the earlier state change persisted. The endpoint must not be chosen solely from the first frame that favors a model. A stronger evaluation would include situations in which earlier actions matter after temporary occlusion or camera motion.

External validity remains separate from a well-controlled local comparison. Alternative controls tested from familiar histories do not establish correct intervention in every game or under every plugin combination. The experiment should identify which variation is intentionally held out and what remains untested. A counterfactual design can make a narrow claim more credible without supporting an unrestricted statement about a generated world's physics. The boundary is especially important when a public demonstration contains only a small set of curated histories.

The ability to accept commands changes the interpretation of a generated sequence. GameNGen's interactive setting (Valevski et al., 2024) and Genie's action-controllable worlds (Bruce et al., 2024) make this distinction visible. A command-response pair needs to be judged for its consequence as well as its visual continuity. Image structure (Wang et al., 2004), learned perceptual correspondence (Zhang et al., 2018), and video-distribution similarity (Unterthiner et al., 2018) supply useful but incomplete observations. The evaluator should explain whether a reference trajectory exists and whether the same initial condition can be replayed. Without those records, a visually different outcome might reflect a valid stochastic alternative or a failure to obey control. The experiment must distinguish those interpretations through its task specification. Matching output length alone is insufficient, because a longer sequence can contain more opportunities for a delayed or persistent control error. The reported endpoint should remain connected to the history that made the interaction meaningful.

Replacing an expensive judgment with an automated score changes the measurement process. A positive correlation is evidence of co-movement, but it does not establish that the two procedures can be used interchangeably (Bland and Altman, 1986). A calibrated forecast has a different requirement: its stated probabilities should match relevant observed frequencies (Guo et al., 2017). Neither property follows merely from a favorable overall ranking. A useful comparison preserves paired measurements, identifies the conditions in which disagreement matters, and keeps scale transformations explicit. If the intended judgment concerns recognizability, playability, or physical compatibility, the proxy should be assessed

against that judgment directly. Agreement on easy cases cannot supply evidence about the difficult conditions absent from the comparison.

4. Reliable evidence about the interaction process

The distribution of available behavior constrains what can be inferred from recorded interaction. The reinforcement-learning framework distinguishes the information in a trajectory from the policy that produced it (Sutton and Barto, 2018). Offline reinforcement learning makes the limitations of fixed experience explicit (Levine et al., 2020), and conservative Q-learning studies one way to address out-of-distribution action evaluation (Kumar et al., 2020). The corresponding lesson for rendering is methodological: an archive of familiar actions is not a complete intervention experiment. If commands are chosen by an existing policy, some states and responses may remain poorly represented. An evaluation should document this support and avoid treating absent cases as successful extrapolation. Proposed counterfactual tests should compare matched initial histories where feasible and retain whether a valid response was produced. Such evidence can strengthen a limited claim without implying that the system models every possible action or game mechanic.

A system's reliable operating region may be narrower than the collection of inputs an interface accepts. Selective prediction separates performance on accepted cases from the fraction of cases accepted (Geifman and El-Yaniv, 2017). Reporting only the former can hide a substantial loss of coverage. Uncertainty estimates may also change under distribution shift (Ovadia et al., 2019), so familiar-condition confidence is not automatically evidence about new subjects, scenes, or control histories. A useful evaluation records rejected, ambiguous, and unsupported cases alongside successful ones. This makes it possible to distinguish limited but honest capability from a system that answers broadly with unreliable certainty. Any proposed deferral rule still requires assessment on its intended input population; it is not validated merely by being described.

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

The proposed test treats control as a contrast between required consequences under matched histories, not simply a visible change after an input. It retains stochastic variation, delayed effects, and failed measurements while separating action fidelity from appearance. Applying it would require actual generated outputs and an independent task rule. The conceptual contribution is a recoverable design for such evidence. A generative game responds meaningfully when the intended alternative intervention changes the relevant outcome, within a scope the evaluation can actually observe.

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.