Procedural Variation as a Test of Generative Game Transfer

Wesley Sullivan (VE), Jeffrey King (AG), Thomas Gardner (NO)

Abstract

New layouts within a familiar game are different evidence from new game mechanics. This conceptual study proposes a hierarchy of procedural variation for interactive-rendering evaluation, separating changed instances, changed visual styles, and changed task rules. It also specifies how histories and generation settings should be partitioned before model selection. Make a game motivates the interface problem, but no transfer experiment is reported. The proposal supports bounded claims about the variation actually held out and avoids treating a new random seed or visual texture as proof of generalization to a new environment family.

1. A new instance may preserve most of the problem

A procedurally generated environment can change its layout while retaining familiar objects, controls, and task rules. This is a meaningful form of variation, but it is not equivalent to learning a new mechanic. A renderer can succeed on a new arrangement because the important transitions remain familiar. It can also fail after a visual style change even when the underlying task has not changed. An evaluation should identify which properties vary and which remain fixed rather than use novelty as an undifferentiated label.

Make a game motivates studying a generative interface across interactive scenarios (Meng et al., 2026a). The public description does not provide enough detail to certify which scenarios were held out or how their variation was produced. This article therefore develops an evaluation hierarchy as a research design. It does not claim that the proposed renderer has passed or failed any particular procedural test. The objective is to connect the wording of a transfer claim to the information the evaluation actually withholds.

A benchmark's title does not specify every property of its released files. ALE establishes a platform and task population (Bellemare et al., 2012); Gym describes an interaction interface (Brockman et al., 2016); Procgen studies procedural variation (Cobbe et al., 2019). Each addresses a different aspect of evaluation. A secondary analysis of BBF should follow the task labels and scores actually present in the downloaded release rather than infer coverage from the associated paper's headline setting (Schwarzer et al., 2023). Statistical evaluation also benefits from retaining distributions instead of presenting only one preferred aggregate (Agarwal et al., 2021). Applied to generated interaction, this means recording environments, histories, available controls, and repeated outputs explicitly. A reader should be able to tell whether a result concerns familiar states, unseen histories within familiar games, or new environment families. Those are different forms of evidence, even when all are measured through a common software interface and summarized using a similarly named score.

Learned world models connect visual representation to behavior in several ways. World Models explores compressed representations of environment experience (Ha and Schmidhuber, 2018). Dreamer uses latent imagination to support behavior learning (Hafner et al., 2019), IRIS studies transformer-based world modeling (Micheli et al., 2022), and DIAMOND investigates diffusion-based representations in which visual detail can matter to control (Alonso et al., 2024). These works establish a useful set of questions rather than a single interchangeable implementation. What information is retained, which histories condition prediction, and how is a predicted observation connected to an action consequence? A rendering study should state which of these questions it actually tests. Visually plausible outputs alone do not establish an accurate transition model. Conversely, an agent can obtain a favorable return without

reproducing all details relevant to a human viewer. The measurement design should preserve this distinction and avoid translating a result from one endpoint into an unobserved improvement on another.

2. Defining the variation before drawing instances

One level could vary initial positions and layouts while keeping object categories, controls, and transition rules fixed. A second could change appearance while preserving those rules. A third could introduce new combinations of familiar mechanics, and a fourth could alter a mechanic itself. These are proposed design categories rather than a universal taxonomy. Their purpose is to prevent a favorable result at an easier level from being silently described as evidence at a harder one.

For each level, the record should identify the procedure that generates instances and the parameters withheld during development. A seed label is insufficient if the same resulting layouts or nearly identical histories occur on both sides of the split. Conversely, two seeds can generate overlapping conditions legitimately if the intended claim concerns repeated draws from a known population. The split should follow the scientific question, and the release should preserve enough instance information to check the relationship between development and assessment cases.

The endpoint must also remain meaningful across variation. A visual comparison against one reference trajectory can become inappropriate when several continuations are valid. A task rule may provide a more stable criterion for action consequences, accompanied by an appearance assessment that permits the intended diversity. The evaluation should not penalize a correct new layout merely because it differs from a familiar screenshot. Nor should it grant success to an attractive scene that violates the unchanged mechanic under test.

A large collection of measurements does not guarantee a large collection of independent cases. Repeated observations can share unrecorded acquisition conditions or a common source, making row-level uncertainty misleading. White's covariance estimator addresses heteroskedasticity within a specified regression (White, 1980), while cluster methods address a declared form of grouped dependence (Cameron and Miller, 2015). Neither approach repairs a poorly defined outcome or establishes representativeness of the observed sample. A report should identify the number of groups, their relative sizes, and the reason for selecting the grouping level. When only a few independent groups are available, apparently precise numerical intervals warrant particular caution. This is a limitation of the evidence structure, rather than a reason to count correlated measurements as fresh observations.

A proposed explanation should be distinguishable from alternative explanations using observations the study can actually collect. Diagramming the relationships among selection, exposure, measurement, and outcome makes that requirement explicit (Greenland et al., 1999). The diagram does not establish its own correctness, but it prevents adjustment choices from remaining invisible. The same transparency is needed when investigators inspect many endpoints before choosing a narrative. Controlling false discovery concerns the behavior of a specified testing family (Benjamini and Hochberg, 1995); it is not a general certificate of scientific validity. Primary endpoints, secondary descriptions, and future mechanism tests should be named separately. Readers can then assess whether an attractive pattern was predicted, explored, or merely noticed after the available comparisons were examined.

Control and fidelity should be assessed as related but distinct properties. Neural game generation in GameNGen (Valevski et al., 2024) and action-controlled environment generation in Genie (Bruce et al., 2024) bring both into view. Structural similarity (Wang et al., 2004) or a learned image distance (Zhang et al., 2018) may help quantify a reference comparison; a video-distribution metric (Unterthiner et al., 2018) can summarize another aspect of temporal appearance. These measurements do not identify whether the

interface interpreted a command correctly. The evaluation record should link an output to the exact control history and to any available state observation. It should also distinguish permitted variation from a broken requirement. A stochastic rendering system need not reproduce every pixel to be useful, but it does need to preserve the aspects of interaction that its claim depends on. Measuring those requirements explicitly makes the relationship between visual quality and behavior open to examination.

3. Interpreting transfer without hiding its denominator

A proposed study would report the number of generated instances, valid interactions, measurable outputs, and task successes at every declared level. This matters because a system may reject more unfamiliar cases or fail before a score can be computed. Comparing only accepted outputs could make the most restrictive operating region appear strongest. A useful result should retain both performance and coverage, along with the reason that a case was excluded or could not be assessed.

Adaptation opportunities must also be matched. A method that receives new reference histories, fine-tuning, or manual plugin changes has been given information unavailable to a method evaluated without adaptation. Such a comparison may reflect a realistic workflow, but it should not be called the same zero-adaptation task. The resource record should distinguish preparation for an environment from the repeated generation of outputs once that preparation is complete. The evaluation can then compare the workflows without obscuring what each system was allowed to learn.

The strongest conceptual contribution of procedural testing is a controlled ladder of questions. If performance changes only when mechanics change, that suggests a different research direction from a failure under harmless appearance variation. The observations would still need careful follow-up before assigning an internal cause. A hierarchy of tests is valuable because it makes those alternatives visible, not because it automatically explains them. No result on a finite ladder establishes arbitrary environmental competence beyond its observed support.

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.

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

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

Procedural variation can make transfer claims more precise when the changed and unchanged properties are documented. New layouts, appearance changes, new combinations, and new mechanics should be interpreted as different assessment boundaries. The proposed design protects those boundaries, matches adaptation access, and retains coverage alongside outcomes. It is a plan for a direct renderer study rather than a completed experiment. Its value is to make a limited success claim clear enough that a stronger claim can be tested separately.

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.