Response Time and Event Order in Generative Game Interfaces

Owen Stevens (SV), Cole Barker (SA), Christian Reed (ID)

Abstract

Rendering speed does not determine whether an interaction arrives in time to be useful. This conceptual analysis separates command reception, visible response, completion of the intended state change, and the order of competing events. It proposes reporting latency distributions alongside task consequences and visual quality under documented hardware and load. Make a game motivates the interface question, but no new timing experiment is performed. The central argument is that a faster first image can coexist with a delayed, incomplete, or wrongly ordered response, so a single frames-per-second number cannot establish responsive control.

1. The time at which an action matters

An interactive user may issue a command because an object is approaching, a decision window is closing, or a previous action has not yet taken effect. In such situations, responsiveness concerns a relation between events rather than the speed of producing arbitrary images. A system can display many frames while continuing to show an old state. It can also present an early visual cue before completing the consequence the user intended. Both behaviors can look fast under an incomplete timing definition.

Make a game places generated visual content within an interactive control setting (Meng et al., 2026a). Its public author description mentions unified tokens and lightweight plugins, which makes the boundary between command handling and rendering a useful design question. That description does not supply measured end-to-end delays or a verified implementation trace. This article therefore develops an evaluation design rather than attributing timing results to the paper. The proposed observations would be needed before a claim about responsive interaction could be assessed directly.

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.

2. Distinguishing receipt, response, and completion

A timing record should first identify the instant at which the interface receives a command. A second event marks the earliest observable output that is attributable to that command. A third marks completion of the required consequence. These events should not be collapsed automatically. For example, a change in camera orientation may be visible quickly while the controlled object continues along an outdated trajectory. The first response delay would be short even though the relevant task remains poorly served.

The definition of completion must follow the task. Some commands request an immediate discrete change; others initiate a continuing process. A protocol can specify a bounded observation window and a

condition that must hold within it, while recording cases that never satisfy the condition. Those noncompletions should not disappear from a summary of successful response times. A latency distribution conditional on success should be accompanied by its success fraction. Otherwise, an interface that fails on slow or difficult cases can appear unusually fast.

Hardware, load, resolution, model version, and plugin configuration also belong in the record. A network delay, queueing delay, preprocessing stage, and generation stage may contribute differently. Separating these components can explain an observed bottleneck, but the user-facing endpoint should remain end-to-end. A favorable internal kernel time does not establish timely control when the rest of the path is unmeasured. No particular timing decomposition is assumed to describe Make a game's actual implementation without direct evidence.

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.

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.

A validation split defines which form of transfer has been observed. Related observations can share subjects, scenes, prompts, or acquisition conditions, so randomly splitting rows may leave important information on both sides of the boundary. Structured cross-validation addresses these dependencies (Roberts et al., 2017). Model selection introduces another boundary: choosing a representation, threshold, or tuning parameter using the final assessment data makes the resulting error estimate optimistic (Varma and Simon, 2006). The appropriate held-out unit follows the intended future use. Generalization to a new frame, a new trajectory, and a new subject are not equivalent claims. A study should record those units before fitting or choosing a scoring rule, and preserve an assessment set that has not become part of iterative development.

3. Event order and the cost of a rapid mistake

Two commands can compete. A turn followed by a stop may need a different outcome from a stop followed by a turn. A renderer that processes both quickly but reverses their order has a sequencing problem rather than a conventional throughput problem. The test should preserve the command order, timestamps, and intended rule for simultaneous or conflicting events. If the interface intentionally coalesces inputs, that behavior should be part of its specification and assessed as such.

Repeated trials should also distinguish permissible stochastic visual variation from variation in event order. A generator may legitimately produce different textures or small motion details while preserving the same required response. Conversely, nearly identical images can conceal a missed command. An experiment could compare fixed histories under matched command schedules and vary the available computation or load independently. The relevant outcome would be the distribution of task-valid response delays, not merely a count of generated frames.

There may be an explicit quality-delay tradeoff. Lower resolution, shorter context, or reduced sampling can accelerate an output while removing information needed for stable control. The evaluation should report the chosen operating point and preserve both endpoints. An application might prefer a visibly simpler but timely response; another might tolerate delay for higher-quality inspection. A single universal ranking obscures those preferences. A useful report makes the operating region and the associated consequences available for comparison.

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.

An association can suggest a research direction without identifying an intervention. Causal diagrams make it possible to state which variables precede the proposed change, which are consequences, and which selection processes remain uncontrolled (Greenland et al., 1999). Their value lies in exposing assumptions rather than decorating a regression with a causal label. A second issue arises when many outcomes or subgroups are explored: selecting the most favorable comparison can exaggerate the impression of support. False-discovery procedures address a defined family of inferential tests (Benjamini and Hochberg, 1995). They do not turn an exploratory pattern into a replicated mechanism. A careful report distinguishes prespecified comparisons, descriptive sensitivities, and hypotheses requiring new observations. This distinction is especially important when an evaluation offers many plausible metrics.

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. From a task contrast to a controlled study

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.

The unit of provenance should be fine enough to reconstruct an important disagreement. Reusable identifiers can connect a reported score to its input evidence and the relevant model and evaluator versions (Wilkinson et al., 2016). Pipeline dependencies should be treated as part of the experiment because apparently minor changes can alter output eligibility or metric behavior (Sculley et al., 2015). A report that preserves only aggregate numbers prevents this reconstruction. Keeping per-case results and documented transformations enables alternative weighting and error analysis without repeating expensive generation. Such a record also helps identify whether two papers compare the same scientific objects or merely use similar names for different releases. Versioned evidence makes progress easier to assess across successive evaluations.

5. Conclusion

A responsive generative interface must preserve both the timing and the order of required consequences. Measuring only output rate can miss stale states, incomplete commands, and rapid sequencing errors. The proposed record links command reception to visible response and task completion, retains noncompletion, and documents the conditions under which timing is observed. These are requirements for a future direct test, not new performance results. They provide a concrete way to make an interactive-rendering speed claim meaningful to the task it is supposed to support.

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.