Human Judgments of Playability alongside Visual Fidelity

Evan Walsh (CZ), Grant Douglas (IN), Logan Dawson (IS)

Abstract

A visually impressive clip need not provide a usable interaction. This conceptual analysis proposes a human study that separates appearance judgments from task-based playability and records the histories observers receive. It addresses blinding, order, repeated subjects, unsupported controls, and the interpretation of preferences. Make a game supplies the motivating interactive setting. No participants are recruited and no ratings are analyzed here. The design asks what a human evaluation must observe before it can support a claim about responsive, understandable control rather than a preference for selected screenshots.

A viewer and a player receive different evidence

A viewer can assess a clip's appearance without deciding whether an input produced its intended consequence. A player must form expectations about commands, timing, and ongoing state. The two perspectives can disagree without either being irrational. A visually polished continuation may be difficult to control, while a simple rendering can make the task understandable and responsive. A study should therefore identify whose judgment it needs and what evidence that person receives.

Make a game places generated imagery in an interactive interface (Meng et al., 2026a). That motivates an evaluation beyond passive visual preference, but the available public description does not establish a complete participant protocol or user-study result. This article proposes such a design without attributing new measurements to the paper. Its central distinction is between liking the output and being able to use the interaction to achieve a declared task. Either can matter, but neither should stand in for the other without evidence.

Defining the task and the judgment separately

A proposed study could ask participants to complete a small set of controlled interaction tasks, while another assessment collects appearance judgments on corresponding outputs. The playability endpoint should specify what constitutes task completion and what assistance is permitted. The appearance endpoint should state whether observers compare clips, give ratings, or choose between alternatives. Preferences and absolute ratings produce different data and should not be described as interchangeable measures of a single latent quality.

Observers need the information relevant to their assigned judgment. A person asked whether a command was obeyed should see the command and sufficient prior history. A person asked only about visual coherence might receive a different presentation. Giving every rater only the final frame would make some task questions impossible to answer. The protocol should retain presentation order, available controls, instructions, and any explanations supplied by the experimenter. Those details can change how understandable an interface appears.

Blinding and order require explicit planning. Model names, familiar branding, or unequal clip lengths can influence a preference independently of task performance. A design can mask source labels and balance presentation order while preserving the information necessary to understand the task. Repeated observations from the same participant or history should remain linked. Counting each rating as an independent person would inflate the apparent amount of evidence. A study should plan its observation hierarchy before choosing a convenient analysis.

Interpreting a preference for the complete workflow

A workflow includes more than the final clip. Setup effort, wait time, failed attempts, and manual corrections can determine whether a participant finds an interface usable. If a study presents only selected successful outputs, it estimates a judgment conditional on that selection. That may be a useful task, but it is not the same as evaluating a first-use experience. The protocol should state whether participants observe failures and whether they may request additional generations or alter controls.

A method may support fewer tasks yet perform them clearly. Another may expose more flexibility with a higher rate of ambiguity or failure. A human study should preserve that difference through coverage and task-specific outcomes. Asking only which model is better overall can conceal the reason for a preference. Follow-up questions can identify whether a participant valued appearance, response, predictability, or flexibility, provided those questions are prespecified and do not become an unrestricted search for a favorable story.

The study's population also limits its claims. Experienced players, developers, and unfamiliar users may interpret the same controls differently. A result from one recruited group should not be generalized automatically to every intended audience. The report should describe relevant experience and the learning period permitted. It should also distinguish improvement during practice from stable task performance. These are requirements for a future study, not participant characteristics or findings observed in this conceptual article.

A game can be watchable and difficult to control

Raters can carry over expectations from a beautiful preview, changing subsequent reports without changing the interface. The issue becomes concrete in a hypothetical assessment. A participant rates a smooth clip highly but struggles to reproduce the action shown when allowed to play the same scenario. Use paired play session to connect passive viewing judgment and interactive goal-directed session to difference between experienced control and judged visual quality. Randomize the first exposure and keep successful task execution distinct from self-reported satisfaction.

Randomly assigning some participants to watch first and others to play first creates a check on preview effects. Give the same goal, control description and allotted time to both groups. Record whether the goal was met and whether the participant understood which action caused each important event. A visually rich preview can alter confidence about the upcoming interface before any action is taken. Analyse that experience separately from the success rate and preserve the order of exposure in the evaluation record.

A result about difference between experienced control and judged visual quality needs a denominator that includes all eligible attempts, including ones without a usable score. For the paired play session, log interruptions and timeouts separately so readers can see whether a response was observed.

Record attempts in their original order and store state, input and response separately for paired play session. If there are many observations of paired play session, decide which ones share an underlying history and keep that grouping in any uncertainty analysis. For interactive goal-directed session, compare conclusions after varying the permitted response window within a declared range.

A sensitivity display can show where the conclusion about difference between experienced control and judged visual quality changes as assumptions about the eligible observations change. For passive viewing judgment, report whether agreement concerns game progress, visible state or the timing of the response. A transfer study of difference between experienced control and judged visual quality would need fresh cases chosen independently of the observed failures.

Measurement and error

Local structural agreement can be inspected alongside task metrics when pixel changes are expected but geometry or regional coherence should remain stable. For a prospective comparison, scaling choices determine how units of different magnitude contribute to an aggregate error; a transparent report retains the original-scale results where interpretable. An existing method compares choices of forecast error scale across series (Hyndman and Koehler, 2006). A methodological precedent compares the local structure of a pair of images (Wang et al., 2004).

Video feature distributions offer a sequence-level view whose unit differs from the individual frame or event used in a diagnostic review. A paired difference plot can reveal systematic offsets and score-dependent disagreement even when the correlation between two procedures is high. For a prospective comparison, a methodological precedent examines discrepancies between paired measurement procedures (Bland and Altman, 1986). One approach extends distributional evaluation to video features (Unterthiner et al., 2018).

Earlier research organizes video-generation evaluation into distinct measurement dimensions (Huang et al., 2023). In practical terms, learned perceptual resemblance responds to some structural differences that pixel errors miss, while remaining dependent on the crop and the chosen comparison target. Its separate video dimensions provide a reason to publish a failure profile instead of combining every observation into a single quality index. A relevant contribution uses deep features to measure pairwise perceptual resemblance (Zhang et al., 2018).

Systems and comparators

Latent rollout and task reward are different outputs, and their errors can be isolated by comparing predictions over a common initial history. Data-efficiency comparisons benefit from matching environment-interaction budgets and specifying whether any additional planning or pretrained information differs between systems. Earlier research uses imagined trajectories in learned latent dynamics for control (Hafner et al., 2019). Related work investigates data-efficient Atari learning using a learned planning model (Ye et al., 2021).

A learned latent action and a recorded controller action require separate interpretations, especially when assessing whether two identical commands produce comparable state transitions. Cross-domain reporting is most informative when the analysis preserves environment-specific scores rather than treating differently scaled rewards as interchangeable observations. One study uses world models across several control domains (Hafner et al., 2023). A methodological precedent learns interactive environments from video through latent action modeling (Bruce et al., 2024).

First-frame conditioning gives a clear initial anchor; later tests can distinguish ordinary appearance change from drift after a break in visibility. Compact latent state can aid control while discarding visible detail. Evaluate policy utility and image fidelity separately. For a prospective comparison, one approach investigates learned latent world representations for sequential control (Ha and Schmidhuber, 2018). One study uses first-frame information to guide spatial and temporal consistency (Ren et al., 2024).

An ablation can hold histories fixed and alter the predictive model, separating changes to rollout quality from changes in task selection. Standard environment interfaces expose reset and step events that can serve as audit points when observations and actions are exported from different experiments. In an evaluation, a relevant contribution uses a transformer world model for sample-efficient reinforcement learning (Micheli et al., 2022). A relevant contribution provides standard environment interfaces for reinforcement learning research (Brockman et al., 2016).

Perturbing action execution is one way to test reliance on deterministic trajectories while preserving a common command vocabulary across comparisons. An existing method defines sequential decision problems in terms of state, action and return (Sutton and Barto, 2018). Earlier research uses sticky actions and standardized protocols to revisit Atari evaluation (Machado et al., 2017). For implementation, a sequential decision formulation keeps reward, action, observation and state distinct, which helps determine where a claimed capability can actually be measured.

An agent trained from pixels provides a control for task success that can be compared with direct assessments of rendering or transition fidelity. In practical terms, an explicit distinction between termination and truncation is relevant when the outcome depends on the elapsed evaluation window or on incomplete episodes. A relevant contribution learns action values from Atari screen observations (Mnih et al., 2013). One published study specifies interfaces for environments and reinforcement learning experiments (Towers et al., 2024).

An existing method investigates action-conditioned game video and transfer to new scenes (Yu et al., 2025). In a controlled test, the distribution that generated an archive constrains follow-up evaluation, particularly where difficult states or rare recovery actions are sparsely represented. Tests on unseen scenes and seen control types can separate transfer to new appearance from transfer to new dynamics and commands. An existing method reviews the problem of learning policies from a fixed set of experience (Levine et al., 2020).

Earlier research uses a diffusion world model to investigate visual detail in Atari (Alonso et al., 2024). A relevant contribution demonstrates action-conditioned neural generation in a game engine setting (Valevski et al., 2024). Evaluate generated images and agent returns separately, while recording whether competing models received the same interaction histories. Action conditioning lets an evaluator test whether a command changed the game despite the plausibility of the rendered frames.

Samples and evidence

The task and knowledge inventories help separate performance on well-represented activities from evidence about broader transfer to unusual interactions. In a controlled test, an existing method uses procedurally generated environments to study generalization (Cobbe et al., 2019). A relevant contribution organizes open-ended Minecraft tasks and external knowledge resources (Fan et al., 2022). Holding out level generation seeds permits a cleaner transfer test than adding repetitions of environments that the system already encountered.

Analysis and interpretation

Repetitions from a common generating condition can have correlated errors, making the number of independent groups more informative than the nominal row count. During validation, serial dependence makes neighboring measurements count differently for uncertainty, even if the average contrast itself is calculated using all observed time points. An existing method allows covariance estimates to account for heteroskedasticity and serial dependence (Newey and West, 1987). One study provides guidance on inference under dependence within specified groups (Cameron and Miller, 2015). An existing method connects validation splits to temporal, spatial and hierarchical dependence (Roberts et al., 2017).

A proposed causal diagram can distinguish decisions made before a run from variables created by the system and observed only after it responds. During validation, conditioning on a desired return does not establish that the archive contains successful examples of the associated trajectory or command sequence. Related work conditions sequence modeling on returns in offline control (Chen et al., 2021). Another

study uses causal diagrams to articulate assumptions about adjustment and selection (Greenland et al., 1999).

Conclusion

Human evaluation of interactive rendering should give playability and visual fidelity separate, interpretable observations. The proposed design links judgments to task instructions, histories, and the complete workflow, while preserving participant and case dependencies. It avoids interpreting a preference for selected clips as proof of usable control. Direct participant evidence would still be needed. The conceptual contribution is a recoverable protocol for obtaining that evidence and stating which user experience the resulting judgment actually represents.

References

Huang, Ziqi, et al.. "VBench: Comprehensive Benchmark Suite for Video Generative Models" *arXiv preprint arXiv:2311.17982*, 2023. https://arxiv.org/abs/2311.17982.

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.

Yu, Jiwen, et al. "GameFactory: Creating New Games with Generative Interactive Videos." *arXiv preprint arXiv:2501.08325*, 2025. https://arxiv.org/abs/2501.08325.

Valevski, Dani, et al.. "Diffusion Models Are Real-Time Game Engines" *arXiv preprint arXiv:2408.14837*, 2024. https://arxiv.org/abs/2408.14837.

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.

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.

Ye, Weirui, et al. "Mastering Atari Games with Limited Data." *arXiv preprint arXiv:2111.00210*, 2021. https://arxiv.org/abs/2111.00210.

Hafner, Danijar, et al.. "Dream to Control: Learning Behaviors by Latent Imagination" *arXiv preprint arXiv:1912.01603*, 2019. https://arxiv.org/abs/1912.01603.

Hafner, Danijar, et al. "Mastering Diverse Domains through World Models." *arXiv preprint arXiv:2301.04104*, 2023. https://arxiv.org/abs/2301.04104.

Bruce, Jake, et al.. "Genie: Generative Interactive Environments" *arXiv preprint arXiv:2402.15391*, 2024. https://arxiv.org/abs/2402.15391.

Ha, David and Jürgen Schmidhuber. "World Models" *arXiv preprint arXiv:1803.10122*, 2018. https://arxiv.org/abs/1803.10122.

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.

Hyndman, Rob J. and Anne B. Koehler. “Another look at measures of forecast accuracy.” *International Journal of Forecasting*, vol. 22, no. 4, 2006, pp. 679-688. https://doi.org/10.1016/j.ijforecast.2006.03.001.

Ren, Weiming, et al. "ConsistI2V: Enhancing Visual Consistency for Image-to-Video Generation." *arXiv preprint arXiv:2402.04324*, 2024. https://arxiv.org/abs/2402.04324.

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.

Brockman, Greg, et al.. "OpenAI Gym" *arXiv preprint arXiv:1606.01540*, 2016. https://arxiv.org/abs/1606.01540.

Micheli, Vincent, et al.. "Transformers are Sample-Efficient World Models" *arXiv preprint arXiv:2209.00588*, 2022. https://arxiv.org/abs/2209.00588.

Cobbe, Karl, et al.. "Leveraging Procedural Generation to Benchmark Reinforcement Learning" *arXiv preprint arXiv:1912.01588*, 2019. https://arxiv.org/abs/1912.01588.

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.

Sutton, Richard S., and Andrew G. Barto. *Reinforcement Learning: An Introduction*. 2nd ed., MIT Press, 2018. http://incompleteideas.net/book/the-book-2nd.html.

Machado, Marlos C., et al. "Revisiting the Arcade Learning Environment: Evaluation Protocols and Open Problems for General Agents." *arXiv preprint arXiv:1709.06009*, 2017. https://arxiv.org/abs/1709.06009.

Fan, Linxi, et al. "MineDojo: Building Open-Ended Embodied Agents with Internet-Scale Knowledge." *arXiv preprint arXiv:2206.08853*, 2022. https://arxiv.org/abs/2206.08853.

Mnih, Volodymyr, et al. "Playing Atari with Deep Reinforcement Learning." *arXiv preprint arXiv:1312.5602*, 2013. https://arxiv.org/abs/1312.5602.

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.

Towers, Mark, et al. "Gymnasium: A Standard Interface for Reinforcement Learning Environments." *arXiv preprint arXiv:2407.17032*, 2024. https://arxiv.org/abs/2407.17032.

Newey, Whitney K., and Kenneth D. West. “A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix.” *Econometrica*, vol. 55, no. 3, 1987, pp. 703-708. https://doi.org/10.2307/1913610.

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.

Chen, Lili, et al. "Decision Transformer: Reinforcement Learning via Sequence Modeling." *arXiv preprint arXiv:2106.01345*, 2021. https://arxiv.org/abs/2106.01345.

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.

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.