Persistence after Camera Motion in Interactive Visual Worlds

Matthew Fleming (MG), Grant Chapman (GT), Leo Fisher (VA)

Abstract

An object that leaves the image creates a useful test of a generated world's persistence. This conceptual study proposes controlled camera-away-and-return histories, with separate observations of object existence, location, state, and appearance. It distinguishes a camera change from an object change and a permitted stochastic variation from a lost task requirement. Make a game motivates the interactive setting, but no new clips are generated here. The design aims to identify what evidence would support persistence beyond short-range visual continuity and what remains unknowable when the underlying state cannot be observed.

Leaving the image is not leaving the world

A camera can turn away while a controlled object continues to exist elsewhere in the environment. When the camera returns, the evaluation may need to establish whether the object is still present, has an appropriate location, and retains changes caused by earlier actions. A plausible returning view does not automatically satisfy those requirements. The generator might create a visually similar replacement, restore an outdated state, or place the object in a different relation to its surroundings.

Make a game frames rendering as an interactive process (Meng et al., 2026a). That framing makes persistence a useful proposed capability to examine, but the available public description does not establish a particular memory mechanism or a completed persistence test. This article supplies a research design rather than a claim about the paper's implementation. It asks how an evaluator could distinguish a stable world from a sequence of individually convincing views when camera motion temporarily removes the relevant evidence.

A history that can reveal lost state

The proposed protocol would begin with an observable object and a documented state-changing action. The camera would then follow a declared path away from the object and return after a specified interval. The final assessment would compare the returning observation with the task rule, taking account of processes that are allowed to continue while the object is out of view. A moving object should not be required to remain frozen unless that is part of the environment specification.

The record should separate the camera command, the object's commanded action, and any ongoing environmental dynamics. Otherwise, a changed view can be attributed ambiguously to camera geometry, object motion, or a generation error. A useful test might include histories with no object-state intervention and histories with one, while matching other initial evidence. It would also vary the period out of view deliberately rather than infer long-term persistence from a short successful return. These comparisons are proposed observations, not measurements already performed for this article.

The endpoint should be decomposed. Existence asks whether the required object returns at all. Location concerns its relation to a reference frame or task geometry. State concerns whether an earlier interaction remains effective. Appearance concerns whether the returning object is visually compatible with its identity or role. A single visual similarity score may blend these dimensions and reward a familiar-looking replacement. Their definitions should therefore be fixed before a preferred model's outputs are inspected.

What can be established without privileged state access

A conventional environment may provide a state log that makes a persistence claim easier to check. A wholly generative interface may provide only images and commands. In the latter case, the evaluator needs to state which conclusions are observational. It may be possible to test whether later actions behave as if the earlier change persisted, even when the hidden state cannot be inspected. Such a behavioral probe is useful, but it is different from proving that a particular internal object representation was maintained throughout the unseen interval.

Stochastic appearance also requires care. The returning texture need not reproduce every pixel to preserve a task-relevant object. Conversely, stable color and shape do not establish correct location or state. An experiment should distinguish allowed appearance variation from prohibited changes in identity or consequence. Independent human judgments may help interpret ambiguous cases, but they should receive the history needed for the intended judgment rather than only a pair of selected screenshots.

The test population should include more than easy empty scenes. Multiple similar objects, changing illumination, intervening actions, and longer periods away can expose different weaknesses. These conditions should be varied through a declared design rather than assembled retrospectively from failures. A report could provide a persistence profile over the tested durations and conditions, together with measurement coverage. It should not extrapolate that profile to arbitrary time spans or scene complexity. Missing observations remain missing evidence, even when the immediate visual continuation looks convincing.

Inspecting a state after leaving the field of view

A renderer can memorize the last view or infer a plausible room without retaining whether the key was collected. The issue becomes concrete in a hypothetical assessment. A player leaves a room after collecting a key and returns by a route that briefly hides every landmark. Use return-to-view episode to connect return after collecting the key and return after inspecting the key only to consistency of the key state with the earlier event. Combine a changed return path with several event histories that share the same final visible entrance.

The restored room should be probed through a consequence that depends on the earlier event. For example, a door that needs the key tests whether the inventory changed, even if the returning view depicts a plausible empty shelf. A changed route helps prevent a simple last-frame match from solving the test. Additional conditions can hold the final entrance image approximately constant while changing the prior collection history. Score the object and inventory transitions separately so that a plausible appearance receives only its own credit.

An initial table can record the eligibility and observation status of return-to-view episode, before any performance statistic is calculated. For the return-to-view episode, 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 return-to-view episode. When the same design produces several observations of return-to-view episode, start with a group-level contrast and inspect the spread among groups. For return after inspecting the key only, compare conclusions after varying the permitted response window within a declared range.

For consistency of the key state with the earlier event, define one setting for the primary estimate and show the neighboring settings as a separate comparison. For return after collecting the key, report whether agreement concerns game progress, visible state or the timing of the response. A transfer study of

consistency of the key state with the earlier event would need fresh cases chosen independently of the observed failures.

Systems and comparators

An existing method demonstrates action-conditioned neural generation in a game engine setting (Valevski et al., 2024). For implementation, a methodological precedent uses a transformer world model for sample-efficient reinforcement learning (Micheli et al., 2022). Action conditioning lets an evaluator test whether a command changed the game despite the plausibility of the rendered frames. 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. First-frame conditioning gives a clear initial anchor; later tests can distinguish ordinary appearance change from drift after a break in visibility. A methodological precedent uses first-frame information to guide spatial and temporal consistency (Ren et al., 2024). For implementation, a relevant contribution provides standard environment interfaces for reinforcement learning research (Brockman et al., 2016).

Compact latent state can aid control while discarding visible detail. Evaluate policy utility and image fidelity separately. A representation optimized for planning can support decisions without reproducing every visual detail; the relevant distinction can be tested with paired observations. During validation, an existing method investigates learned latent world representations for sequential control (Ha and Schmidhuber, 2018). One published study learns a task-oriented model for planning without supplied game dynamics (Schrittwieser et al., 2019).

An existing method investigates action-conditioned game video and transfer to new scenes (Yu et al., 2025). In a controlled test, one published study defines sequential decision problems in terms of state, action and return (Sutton and Barto, 2018). Tests on unseen scenes and seen control types can separate transfer to new appearance from transfer to new dynamics and commands. A sequential decision formulation keeps reward, action, observation and state distinct, which helps determine where a claimed capability can actually be measured.

The emulator protocol keeps the task definition observable, including how action sets and episode boundaries affect which returns can legitimately be compared. In practical terms, evaluate generated images and agent returns separately, while recording whether competing models received the same interaction histories. Related work defines an emulator-based environment for testing general game agents (Bellemare et al., 2012). A methodological precedent uses a diffusion world model to investigate visual detail in Atari (Alonso et al., 2024).

Data-efficiency comparisons benefit from matching environment-interaction budgets and specifying whether any additional planning or pretrained information differs between systems. One approach learns interactive environments from video through latent action modeling (Bruce et al., 2024). A learned latent action and a recorded controller action require separate interpretations, especially when assessing whether two identical commands produce comparable state transitions. Another study investigates data-efficient Atari learning using a learned planning model (Ye et al., 2021).

Cross-domain reporting is most informative when the analysis preserves environment-specific scores rather than treating differently scaled rewards as interchangeable observations. The distribution that generated an archive constrains follow-up evaluation, particularly where difficult states or rare recovery actions are sparsely represented. During validation, earlier research uses world models across several

control domains (Hafner et al., 2023). One published study reviews the problem of learning policies from a fixed set of experience (Levine et al., 2020).

Performance outside the observed behavior distribution should be reported separately because offline returns depend on which action sequences the training archive supports. Perturbing action execution is one way to test reliance on deterministic trajectories while preserving a common command vocabulary across comparisons. For a prospective comparison, one study uses sticky actions and standardized protocols to revisit Atari evaluation (Machado et al., 2017). A methodological precedent uses conservative value estimation for offline reinforcement learning (Kumar et al., 2020).

Analysis and interpretation

Conditioning on a desired return does not establish that the archive contains successful examples of the associated trajectory or command sequence. A relevant contribution allows covariance estimates to account for heteroskedasticity and serial dependence (Newey and West, 1987). One approach conditions sequence modeling on returns in offline control (Chen et al., 2021). In practical terms, serial dependence makes neighboring measurements count differently for uncertainty, even if the average contrast itself is calculated using all observed time points.

A simulator can vary visibility, traffic and camera placement independently while leaving the underlying events observable for verification against the rendered frames. In an evaluation, a methodological precedent reviews matching as a design step for comparing observational units (Stuart, 2010). Matching can improve comparability on recorded attributes; an overlap display should identify which observations are lost before the outcome contrast is computed. Earlier research provides a configurable driving simulation environment (Dosovitskiy 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. In a controlled test, because several changes coexist in one system, isolating a component requires an intervention or a comparison with the other choices held fixed. A relevant contribution combines several extensions of deep Q-learning in a common Atari agent (Hessel et al., 2017). Related work uses causal diagrams to articulate assumptions about adjustment and selection (Greenland et al., 1999).

Samples and evidence

Distinct behavior-policy mixtures allow an assessment of sensitivity to data support without silently treating all offline experience as drawn from one policy. The task and knowledge inventories help separate performance on well-represented activities from evidence about broader transfer to unusual interactions. During validation, earlier research defines offline reinforcement learning tasks with varied data collection policies (Fu et al., 2020). One approach organizes open-ended Minecraft tasks and external knowledge resources (Fan et al., 2022).

Measurement and error

Earlier research organizes video-generation evaluation into distinct measurement dimensions (Huang et al., 2023). Local structural agreement can be inspected alongside task metrics when pixel changes are expected but geometry or regional coherence should remain stable. In a controlled test, its separate video dimensions provide a reason to publish a failure profile instead of combining every observation into a single quality index. Earlier research compares the local structure of a pair of images (Wang et al., 2004). One study connects validation splits to temporal, spatial and hierarchical dependence (Roberts et al., 2017).

Report which raters, measurements and agreement definition enter the coefficient, since different designs give superficially similar statistics different meanings. Earlier research examines uncertainty estimates under dataset shift (Ovadia et al., 2019). A relevant contribution distinguishes forms of intraclass correlation according to the rating design (Koo and Li, 2016). Confidence calibrated on the original collection may change under a new capture or generation regime, so calibration belongs in transfer testing.

Conclusion

Camera motion provides a controlled way to ask whether a generated interaction preserves more than the current image. The proposed design links earlier state changes to later observations and separates existence, location, state, and appearance. It also identifies the limit of image-only evidence about hidden representations. A future renderer study could use this protocol to test a bounded persistence claim. Without such observations, plausible returning views remain suggestive demonstrations rather than established evidence of a stable interactive world.

References

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.

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

Schrittwieser, Julian, et al. "Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model." *arXiv preprint arXiv:1911.08265*, 2019. https://arxiv.org/abs/1911.08265.

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

Stuart, Elizabeth A. "Matching Methods for Causal Inference: A Review and a Look Forward." *Statistical Science*, vol. 25, no. 1, 2010, pp. 1-21. https://doi.org/10.1214/09-sts313.

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

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.

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.

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

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

Dosovitskiy, Alexey, et al. "CARLA: An Open Urban Driving Simulator." *arXiv preprint arXiv:1711.03938*, 2017. https://arxiv.org/abs/1711.03938.

Fu, Justin, et al. "D4RL: Datasets for Deep Data-Driven Reinforcement Learning." *arXiv preprint arXiv:2004.07219*, 2020. https://arxiv.org/abs/2004.07219.

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.

Hessel, Matteo, et al. "Rainbow: Combining Improvements in Deep Reinforcement Learning." *arXiv preprint arXiv:1710.02298*, 2017. https://arxiv.org/abs/1710.02298.

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.

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.

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.

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.

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.

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

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.

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

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.

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

Koo, Terry K., and Mae Y. Li. "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research." *Journal of Chiropractic Medicine*, vol. 15, no. 2, 2016, pp. 155-163. https://doi.org/10.1016/j.jcm.2016.02.012.

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.

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

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.

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