Long-Sequence Drift in Subject-Preserving Video

Justin Fisher (SK), Oliver Collins (SM), Derek Porter (MN)

Abstract

A strong beginning and an acceptable average can coexist with a subject that changes later in a clip. This conceptual study proposes a time-resolved evaluation of identity drift, distinguishing gradual change, brief substitution, and loss of measurement coverage. ARGUS motivates the conditioning problem, but no long sequence is generated or scored here. The design retains reference access, continuation procedures, and per-window outcomes. It supports a bounded duration claim while avoiding the inference that success on a short clip establishes identity preservation over arbitrary extensions or repeated continuation calls.

1. Duration changes the opportunity for failure

A longer sequence contains more events and more opportunities for a preservation requirement to be violated. It may also contain more varied views and less continuously visible evidence. An evaluation should therefore state whether duration changes only the observation window or also the generation procedure. Extending a single invocation and repeatedly continuing from previous outputs can create different information and error processes. Calling both long-video generation does not make them the same experiment.

ARGUS addresses preservation using structured identity conditioning (Meng et al., 2026b). That motivates asking how its intended requirement should be assessed across time, without assuming a particular unobserved drift behavior. The present article is a conceptual design, not a report of long-run failure rates. No trajectories or internal states are measured. Its contribution is to identify the records and endpoints needed before a duration-dependent preservation claim can be interpreted.

Identity evidence has a sampling structure before any recognition score is calculated. LFW defines a particular pair-verification protocol (Huang et al., 2008), and the documented scikit-learn loader makes its prescribed files accessible through a reproducible software interface (Pedregosa et al., 2011). Neither source turns a pair count into a count of independent subjects or images. Dataset documentation should identify how references, identities, comparisons, and splits were assembled (Gebru et al., 2021). For generated video, the analogous record should distinguish a new person from another reference photograph, another trajectory, or another frame. Each adds evidence along a different dimension. A transparent report should preserve those links rather than flattening every observation into one score table. The resulting metadata can support a useful audit even when no images or generated clips are available. Such an audit describes coverage and organization; it does not itself establish recognition accuracy, demographic representativeness, or identity preservation in an unobserved video.

An identity score is produced by a measurement system with its own training and assumptions. FaceNet (Schroff et al., 2015), ArcFace (Deng et al., 2018), and CosFace (Wang et al., 2018) offer influential approaches to discriminative face representation. Their outputs should be interpreted in relation to the images and decisions on which they are assessed. VGGFace2 broadens variation within subject observations (Cao et al., 2017), MegaFace makes large distractor populations important (Kemelmacher-Shlizerman et al., 2015), and FairFace addresses imbalance across annotated demographic categories (Kärkkäinen and Joo, 2019). These are complementary evaluation concerns. A generated-video study should not compress them into an unexplained claim that a face benchmark is difficult. It should specify whether difficulty comes from viewpoint, acquisition, distractors, subject population, or the representation itself. Demographic conclusions require appropriate annotations and cannot be inferred

merely from identity frequencies or from a generic face-similarity score. A metric's usefulness for one purpose remains an empirical question for another.

2. A time-resolved outcome instead of a single ending

A proposed study would specify observation windows across the full intended duration and retain the same reference identity definition. It would report local judgments, their availability, and the rule linking them to a clip-level result. A final-frame score can miss an earlier substitution, while a simple average can dilute a brief consequential event. A worst-case endpoint can emphasize a single unreliable measurement. The study should choose the rule according to the intended consequence and preserve the data needed to examine alternatives.

Different temporal patterns should remain distinguishable. Gradual drift concerns a sustained change in compatibility with the intended subject. A brief substitution concerns a local event that may later disappear. Lost visibility or measurement failure concerns unavailable evidence. These are proposed outcome categories whose labels would need validation in an actual study. They should not be assigned automatically from a declining numeric score without checking what the instrument measures under changing pose, expression, and occlusion.

Reference access during continuation is part of the task. A workflow may reuse the original reference, add new observations, select frames from its own output, or permit manual correction. Those choices can support useful applications, but they change the information budget. A report should identify them separately and avoid describing a corrected continuation as an untouched long-horizon preservation result. The evaluator also needs to know whether a new generation invocation resets any part of the conditioning process.

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.

The scientific interpretation should specify what would change under the proposed intervention. Holding an observed covariate fixed statistically is not equivalent to manipulating it in an environment. A causal graph helps separate confounders, mediators, and selection mechanisms (Greenland et al., 1999). It can also reveal why controlling for every recorded variable is not automatically defensible. Meanwhile, an evaluation with multiple conditions and endpoints creates a family of opportunities for apparently striking findings. Multiplicity methods require a declared set of comparisons and an appropriate inferential purpose (Benjamini and Hochberg, 1995). Descriptive tables may be useful without testing every cell. When the data do not identify a mechanism, the appropriate contribution can be a sharper question and a transparent comparison rather than an unsupported causal answer.

Subject conditioning is not identical to measuring the resulting subject. DreamBooth studies subject-driven generation through model adaptation (Ruiz et al., 2022), whereas IP-Adapter develops image prompting compatible with text conditioning (Ye et al., 2023). The reference evidence available to those mechanisms should be matched or explicitly described in a comparison. Generic representations

also have different training purposes: CLIP connects imagery to language (Radford et al., 2021), while DINO investigates self-supervised visual features (Caron et al., 2021). Either can contribute to a measurement pipeline, but neither automatically defines human identity across generated motion. A study should separate input fidelity, recognizability, and obedience to scene instructions. It should also retain cases in which these objectives conflict. An image-conditioning mechanism can reproduce an accessory or background very well while leaving a more important identity requirement unresolved. Measuring those conflicts directly is more informative than declaring every increase in reference similarity a successful preservation of the person.

3. Comparing durations without changing the population silently

A longer request may produce fewer valid clips or more unassessable intervals. Comparing only successful complete sequences can favor the longest condition through selection. The study should retain every scheduled request, the amount of output produced, and the reason a measurement is missing. Conditional identity scores should be accompanied by their coverage. A method that supports a shorter reliable operating region can still provide useful evidence if that region is reported honestly.

Duration can also interact with the task. An almost static face and a moving subject traversing several scenes impose different demands. A duration comparison should match or document motion, camera changes, scene transitions, and other requested conditions. Otherwise, the observed difference may concern a changed task mixture rather than temporal persistence alone. A controlled design can nominate a small set of meaningful histories and then test their extensions through a declared procedure.

A useful next step would assess new histories after any drift-driven model repair. Repeatedly choosing thresholds or reference policies to improve the same diagnostic clips turns those clips into development information. An independent assessment boundary is needed for a broader improvement claim. The present proposal does not establish that such a repair succeeds. It provides a way to keep the duration, conditioning, and assessment questions visible throughout that process.

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.

A score is useful only relative to the question it is supposed to answer. Correlation measures a relationship between varying quantities, whereas agreement concerns the discrepancies produced when one procedure substitutes for another (Bland and Altman, 1986). Confidence estimates add a further question about calibration (Guo et al., 2017). These distinctions matter when a learned evaluator becomes the principal source of evidence about another learned system. The evaluation should preserve examples with conflicting scores and obtain independent target judgments where feasible. Numerical ranges and normalization choices should also be documented. An apparently small difference may have a different practical meaning near a decision threshold than in an unimportant part of the scale. Reporting the intended consequence gives those comparisons an interpretable purpose.

4. Connecting an identity claim to its observations

A moving subject introduces quality dimensions that a still-image identity score cannot exhaust. VBench decomposes video-generation evaluation into multiple aspects (Huang et al., 2023). Structural similarity (Wang et al., 2004), learned perceptual correspondence (Zhang et al., 2018), and video-distribution metrics (Unterthiner et al., 2018) provide additional observations under their respective assumptions. They should not be treated as equivalent measurements of the same latent property. A generated clip can have coherent motion and attractive appearance while changing its subject; it can also preserve recognizable identity while containing visible temporal artifacts. Reporting separate outcomes allows those tradeoffs to be examined. The assessment should specify the frames, viewpoints, occlusion states, and durations included in each score. A difficult case that defeats the identity evaluator should remain part of the coverage record rather than disappearing silently. The resulting profile can support a bounded claim about the observed conditions without implying that every aspect of subject consistency has been solved.

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.

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

Identity drift should be assessed through time-indexed observations and an explicit account of continuation and reference access. Averages, ending scores, and worst-case summaries answer different questions, especially when measurement availability changes. The proposed design supports a limited duration claim rather than unlimited persistence. It reports no new ARGUS result. Its value is to make long-sequence evidence interpretable without allowing an attractive start or a favorable total to conceal what happened later in the trajectory.

References

Meng, Zijie, et al. "Argus: Stacked multi-view identity mosaic injection for subject-preserving video generation." *arXiv preprint arXiv:2606.11670* (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.

Huang, Gary B., Marwan Mattar, Tamara Berg, and Erik Learned-Miller. "Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments." *ECCV Workshop on Faces in Real-Life Images*, 2008. https://cs.umass.edu/~elm/papers/Huang_eccv2008-lfw.pdf.

Schroff, Florian, et al.. "FaceNet: A Unified Embedding for Face Recognition and Clustering" *arXiv preprint arXiv:1503.03832*, 2015. https://arxiv.org/abs/1503.03832.

Deng, Jiankang, et al.. "ArcFace: Additive Angular Margin Loss for Deep Face Recognition" *arXiv preprint arXiv:1801.07698*, 2018. https://arxiv.org/abs/1801.07698.

Wang, Hao, et al.. "CosFace: Large Margin Cosine Loss for Deep Face Recognition" *arXiv preprint arXiv:1801.09414*, 2018. https://arxiv.org/abs/1801.09414.

Cao, Qiong, et al.. "VGGFace2: A dataset for recognising faces across pose and age" *arXiv preprint arXiv:1710.08092*, 2017. https://arxiv.org/abs/1710.08092.

Kemelmacher-Shlizerman, Ira, et al.. "The MegaFace Benchmark: 1 Million Faces for Recognition at Scale" *arXiv preprint arXiv:1512.00596*, 2015. https://arxiv.org/abs/1512.00596.

Kärkkäinen, Kimmo and Jungseock Joo. "FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age" *arXiv preprint arXiv:1908.04913*, 2019. https://arxiv.org/abs/1908.04913.

Ruiz, Nataniel, et al.. "DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation" *arXiv preprint arXiv:2208.12242*, 2022. https://arxiv.org/abs/2208.12242.

Ye, Hu, et al.. "IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models" *arXiv preprint arXiv:2308.06721*, 2023. https://arxiv.org/abs/2308.06721.

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

Caron, Mathilde, et al.. "Emerging Properties in Self-Supervised Vision Transformers" *arXiv preprint arXiv:2104.14294*, 2021. https://arxiv.org/abs/2104.14294.

Radford, Alec, et al.. "Learning Transferable Visual Models From Natural Language Supervision" *arXiv preprint arXiv:2103.00020*, 2021. https://arxiv.org/abs/2103.00020.

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.

Gebru, Timnit, et al.. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.

Pedregosa, Fabian, et al. "Scikit-learn: Machine Learning in Python." *Journal of Machine Learning Research*, vol. 12, 2011, pp. 2825-2830. https://jmlr.org/papers/v12/pedregosa11a.html.