Provenance for Reused Identity References and Generated Clips

Peter Clark (CD), Oliver Johnston (TM), Ian Baker (CD)

Abstract

The same reference can pass through cropping, selection, conditioning, generation, and later reuse. Without linked records, a displayed result may no longer reveal which observations produced it. This article proposes a provenance structure for subject-preserving video studies, using ARGUS as the motivating setting. No collection is newly audited. The structure connects source observations, transformations, attempts, and final clips while distinguishing reproducibility metadata from unrestricted access to personal imagery. Its purpose is to make a reported result traceable across a changing research workflow.

1. A result needs a recoverable history

A final clip often appears beside a few reference images, yet those images may differ from the files actually supplied during generation. A crop can change, a selection process can choose another view, or a later export can replace the displayed reference. Such discrepancies need not involve misconduct; they can arise from ordinary workflow evolution. They still matter scientifically because the reader may otherwise attribute an output to reference evidence that was never part of its generation.

ARGUS uses organized multi-view identity information (Meng et al., 2026b). That setting motivates provenance links among the source observations, their conditioning form, and the resulting video. This article proposes a record design and performs no audit of the focal method or any existing collection. The central question is how to preserve a sufficiently precise history for verification while recognizing that a reproducible identifier and unrestricted redistribution of an image are different forms of access.

A benchmark's observation count should be accompanied by a description of what an observation is. The LFW pair protocol separates identities from their images and from the comparisons used for verification (Huang et al., 2008). Its documented software access routes also distinguish the ten-fold assessment from development subsets (Pedregosa et al., 2011). Dataset documentation connects those files to their collection and intended purpose (Gebru et al., 2021). A generation benchmark needs an equally explicit account of how a referenced subject becomes a set of output clips and scored frames. Without it, repeated observations may receive unintended weight or cross a supposed assessment boundary. A metadata analysis can establish whether those links are consistent. It should still avoid claiming that consistency of filenames proves recognizability, fair representation, or robustness under viewpoint change. Those are further outcomes requiring appropriate measurements.

2. Linking sources, transformations, and attempts

A proposed record would assign stable identifiers to source observations, processed derivatives, reference sets, generation requests, attempts, and exported clips. Each derivative would point to its parent observation and recorded transformation. File hashes can help identify exact artifacts, while semantic identifiers preserve relationships across format conversions. Both are useful, but neither alone explains what visual information a model received. The transformation parameters and input versions remain necessary parts of the scientific record.

Generation metadata would include model and software versions, sampling settings, instructions, available references, and the selection process for displayed outputs. Every attempt would retain a status rather than only successful clips entering the record. If a human selects a result, the selection criteria and

candidate pool should be documented. This connects the final output to the actual search opportunity and prevents a curated illustration from being misread as a typical unselected generation under the same request.

Reuse requires another explicit link. A reference or generated clip used in a later experiment should preserve its original identifier and record the new purpose. Renaming the file should not make it appear to be an independent observation. The same principle applies to duplicate publications of a source image. Where duplication cannot be resolved, the uncertainty should remain documented. A provenance graph can describe known relationships without claiming complete knowledge of every earlier appearance of the content.

The choice of statistical unit should precede the choice of a standard error. A dataset can contain many rows but relatively few independently acquired histories. Heteroskedasticity-consistent covariance protects an otherwise specified regression against unequal residual variances (White, 1980). It does not automatically accommodate observations sharing an environment, subject, or recording. Cluster-robust inference makes that dependence part of the variance calculation (Cameron and Miller, 2015). The grouping must follow the observation process and the claim being made; an identifier that is convenient to store is not necessarily the relevant independent unit. For a descriptive analysis, preserving the groups also allows readers to recalculate a different population average. An uncertainty interval remains conditional on the observations and assumptions actually available.

Discriminative identity representations address a different question from the design of the test population. FaceNet learns an embedding for verification and related recognition tasks (Schroff et al., 2015). ArcFace (Deng et al., 2018) and CosFace (Wang et al., 2018) develop margin-based approaches to discriminative face learning. These methods do not determine which reference views should be supplied to a generator. Dataset choices also alter the challenge: VGGFace2 emphasizes pose and age variation (Cao et al., 2017), MegaFace increases the scale of recognition distractors (Kemelmacher-Shlizerman et al., 2015), and FairFace studies representation across annotated demographic groups (Kärkkäinen and Joo, 2019). Their relevance is to separate distinct kinds of coverage. Many comparisons of a small reference set cannot replace missing pose variation or a different population of people. A study should state which identity measurement it uses, how thresholds are selected, and what the represented cases actually cover. Strong performance on one protocol does not establish universal validity of the embedding across generated or stylized subjects.

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.

3. What provenance enables and what it cannot establish

Traceable records support checks of input consistency, selection, and overlap across splits. They can reveal whether two apparently separate experiments share a source observation. They cannot establish identity accuracy or permission merely by existing. Those questions require appropriate additional evidence and task-specific interpretation. The proposed system therefore separates provenance facts from outcome labels, access conditions, and any claims that depend on a broader review beyond the observations recorded in the experiment.

Access can be tiered while preserving useful scientific metadata. A public record may describe identifiers, transformations, and aggregate results, while source imagery remains available only under an appropriate process or may not be redistributable. This arrangement limits some forms of reproduction, and the report should state which ones. A hash does not reconstruct an absent image. Transparency means explaining the boundary rather than promising complete reproducibility from metadata that supports only a subset of the relevant checks.

Versioning is essential when references, clips, or outcome annotations change. A corrected source link can alter an overlap audit, and a revised clip export can change the assessed content. The original result should remain associated with its original artifacts, while the correction receives a new record. This permits scientific improvements without erasing the conditions of earlier comparisons. It also helps distinguish changes in model performance from changes in the evidence collection and processing that surround the model.

Different reference-conditioning mechanisms can receive different amounts of information. DreamBooth adapts a generation model around a subject (Ruiz et al., 2022), while IP-Adapter offers a compatible image-conditioning route (Ye et al., 2023). A fair comparison should preserve how many views, which resolutions, and which adaptation opportunities are supplied. CLIP's language-supervised visual representation (Radford et al., 2021) and DINO's self-supervised features (Caron et al., 2021) can provide useful descriptions of outputs, but their training objectives do not establish a person-recognition standard. The endpoint should specify whether success means matching reference appearance, retaining a recognizable subject, or satisfying a requested scene. Those properties can disagree. A useful analysis would keep the disagreement visible and ask which source of evidence each method uses. The relevant advantage is a measured improvement under a declared information budget, not an unqualified increase in similarity to whatever reference happened to be available.

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.

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. Scope, uncertainty, and independent assessment

An identity evaluation becomes more informative when it retains the other properties of a generated clip. VBench's multiple dimensions (Huang et al., 2023) help motivate this decomposition. Structural similarity (Wang et al., 2004) compares a particular aspect of image structure; learned perceptual similarity (Zhang et al., 2018) uses another representation; video-distribution metrics (Unterthiner et al., 2018) incorporate temporal visual features. None alone establishes a stable person through every pose and event. A study should identify the relevant time windows and examine local failures as well as sequence averages. It should also distinguish an identity evaluator's inability to score a frame from the generator's inability to preserve a subject. Those are different observations with different remedies. A measurement profile linked to the original clip permits this distinction to be checked and avoids attributing broad capability to a favorable average obtained only on easily measurable views.

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

Provenance should connect source observations to derivatives, conditioning sets, attempts, and final clips. The proposed structure reports no new audit findings. It makes future results traceable while keeping outcome validity and access questions distinct from artifact identity. A subject-preservation claim is easier to assess when readers can determine which references and selection opportunities produced it, and which later verification steps the retained records actually make possible.

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.