How Row-Level Repartitioning Breaks an Identity-Disjoint Evaluation

Nathaniel Wallace (PE), Andrew Anderson (UG), Gary Perry (BB)

Abstract

The prescribed LFW folds share no identities across their boundaries. Replacing that organization with random pair-row holdouts changes the assessment substantially. Across 200 seeded repartitions with 600 test pairs, an average of 89.79 percent of test pairs contain at least one identity already present in training; 81.14 percent contain at least one reused image. These are exposure measurements, not observed accuracy inflation. The exercise shows why a convenient row split can answer a different question and motivates preserving subject-level boundaries in evaluations of identity-preserving video.

1. A random split can remove the intended novelty

Randomization of rows does not necessarily create a valid assessment of unfamiliar identities. A pair table contains shared ingredients: people and photographs can appear in several comparisons. Assigning rows independently can therefore place those ingredients on both sides of the split. A model may then be assessed on a different form of transfer from the one implied by a claim about new subjects. The problem is not randomness itself, but a mismatch between the randomized unit and the intended assessment unit.

ARGUS motivates evaluating preservation across subject evidence and generated views (Meng et al., 2026b). That evaluation can require a clear distinction between familiar subjects and unfamiliar ones. The current study does not train or test ARGUS. It performs a computational repartitioning exercise on LFW pair metadata to quantify how an alternative split changes exposure. Its outcome is the presence of shared identity or image keys, not a recognition score, generated-video quality measure, or estimated causal change in accuracy.

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.

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.

2. The resampling exercise actually executed

The original file contains 6,000 pair rows in ten prescribed folds. Direct set checks show no identity or image shared between any original fold and the other nine. For the alternative exercise, the full rows are randomly permuted using a fixed random generator initialized with seed 488. The first 600 rows become a test set and the remaining 5,400 a training set. This operation is repeated 200 times. It is not stratified to force exactly 300 matched and 300 mismatched pairs in each random holdout.

For each repetition, the analysis asks whether a test pair contains at least one identity seen anywhere in the training rows. A second indicator asks whether at least one of its exact image keys appears in training. The average identity-exposed share is 0.8979, with observed repetition values from 0.8567 to 0.9283. The average image-exposed share is 0.8114, with values from 0.7683 to 0.8533. The ranges describe these 200 related assignments; they are not confidence intervals for a classifier's performance.

A face embedding is an instrument whose output needs a defined task. FaceNet supports verification and recognition through a learned metric space (Schroff et al., 2015), while ArcFace (Deng et al., 2018) and CosFace (Wang et al., 2018) pursue discriminative margins through different formulations. The scientific population is supplied by the benchmark, not by those objectives. VGGFace2's within-person variation (Cao et al., 2017), MegaFace's distractor scale (Kemelmacher-Shlizerman et al., 2015), and FairFace's representation emphasis (Kärkkäinen and Joo, 2019) show why populations can differ materially. For subject-preserving video, reference access and output conditions should therefore be documented alongside the identity score. An automated evaluator may be accurate on its original photographic population yet behave differently on generated motion, unusual lighting, or stylization. Establishing the relevant transfer requires observations in those settings. Architectural sophistication in the identity instrument cannot replace a direct assessment of the claim the generator is making.

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.

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. Exposure is evidence about the split, not the model

The contrast with the original zero-overlap organization is large and directly attributable to how pair rows were assigned in this exercise. It does not establish the amount by which a particular model's accuracy would change. A system might exploit shared identity information strongly, weakly, or not at all. Those possibilities require fitted models and predictions under explicitly matched protocols. The metadata analysis instead establishes that the novelty condition has changed before any model is fitted.

The criterion is deliberately at least one shared ingredient. It should not be paraphrased as both identities being known or as the exact pair having been seen. The full file contains no duplicate unordered pair rows, so the exposure arises through repeated people or images across different comparisons. Keeping those definitions precise prevents an exaggerated claim that the same test row was copied into training. The actual design concern is a more subtle overlap in the objects the rows use.

For a generated-subject benchmark, a similar issue can arise when clips, frames, or reference-image combinations are randomized without preserving subject boundaries. A study aimed at new views of familiar people may legitimately allow some shared identity information. A study aimed at unfamiliar people needs a stronger separation. Both can be meaningful, provided their names and input access match the actual protocol. The split should be selected for that purpose rather than inherited from a generic row-shuffling utility.

The exercise also suggests a reproducible preventive check. Before training, the evaluator can publish intersections of subject, reference, and clip identifiers across partitions. Those checks should be tied to the intended boundary rather than require every possible source to be disjoint in every task. A shared scene can be appropriate in one comparison and a confound in another. What matters is that the scientific transfer claim remains recoverable from the assignment record and that development choices do not silently weaken it.

The reference path and the scoring path should be distinguishable in an experiment. DreamBooth (Ruiz et al., 2022) and IP-Adapter (Ye et al., 2023) illustrate different ways to make subject or image evidence influence generation. CLIP (Radford et al., 2021) and DINO (Caron et al., 2021) illustrate representations that can be used to compare visual content for other purposes. Reusing the same representation for conditioning and evaluation may create shared sensitivities that deserve examination. It is not proof of invalidity, but it strengthens the case for an independent human or task endpoint. A comparison should report which information enters each path and retain reference-specific artifacts as potential confounds. The generator's ability to reproduce a background, camera style, or distinctive clothing can be useful while remaining different from preserving identity under a requested change. A careful evaluation makes that difference observable.

Replacing an expensive judgment with an automated score changes the measurement process. A positive correlation is evidence of co-movement, but it does not establish that the two procedures can be used interchangeably (Bland and Altman, 1986). A calibrated forecast has a different requirement: its stated probabilities should match relevant observed frequencies (Guo et al., 2017). Neither property follows merely from a favorable overall ranking. A useful comparison preserves paired measurements, identifies the conditions in which disagreement matters, and keeps scale transformations explicit. If the intended judgment concerns recognizability, playability, or physical compatibility, the proxy should be assessed against that judgment directly. Agreement on easy cases cannot supply evidence about the difficult conditions absent from the comparison.

4. Designing an interpretable preservation study

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.

Difficult cases can disappear at several stages: data eligibility, generation, measurement, and final scoring. Each exclusion changes the population represented by the reported result. Coverage-aware evaluation makes those decisions visible (Geifman and El-Yaniv, 2017). A confidence score should also be checked when the input distribution changes, because uncertainty under shift is an empirical property rather than a guarantee of the model interface (Ovadia et al., 2019). An informative record preserves failures to produce an output and failures to obtain a valid measurement. Treating either as absent data without explanation can inflate apparent reliability. A proposed method should state its unsupported conditions and test the boundary between accepted and rejected cases as carefully as the quality of its successful outputs.

Reproducibility begins with preserving which scientific objects are being linked. Stable identifiers and reusable metadata help connect inputs, reference observations, model versions, and reported outcomes (Wilkinson et al., 2016). The scoring pipeline also has dependencies that can change the interpretation of a result: preprocessing, representation checkpoints, thresholds, and aggregation code can all matter (Sculley et al., 2015). A compact release should retain those decisions rather than only the final mean. Where source data cannot be redistributed, an access description and hashes can still identify the analyzed inputs. This record allows later users to distinguish a change in the underlying system from a change in how the system was measured. Reproducibility is therefore part of interpretation, not simply a final packaging step.

5. Conclusion

Random pair-row holdouts create extensive identity and image exposure across partitions that the prescribed LFW folds avoid. The result concerns protocol organization, not a newly observed change in model accuracy. Subject-preserving video evaluations should therefore state what is intended to be unfamiliar and verify the corresponding identifier boundary before fitting or scoring. A random split can be reproducible while still answering the wrong transfer question. Preserving the intended unit makes its result interpretable.

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.