# Pair Connectivity and the Limits of Transitive Identity Evidence

Owen Hall (CO), Oliver Fleming (SI), Robert Wells (KM)

## Abstract

The positive comparisons in LFW form 1,878 connected image components across 4,720 photographs, although they involve 1,532 identity labels. The largest component contains nine images, and none of the full file's 6,000 unordered pairs is duplicated. These counts distinguish a person label from the subset of image relationships explicitly observed in the protocol. This graph audit does not create new training pairs or recognition outcomes. For subject-preserving generation motivated by ARGUS, it shows why reference connectivity, reference breadth, and measured identity fidelity should remain separate concepts.

## 1. A label does not make every image relation observed

If two photographs share a person label, that metadata can suggest a relation between them. A restricted verification protocol may nevertheless expose only selected pairs. The distinction matters because the protocol defines which evidence a method is permitted to use and which comparisons are actually assessed. Treating every possible relation within a label as an observed pair would change the task. It would also blur the difference between annotation knowledge and measured similarity.

ARGUS uses structured reference observations to support subject-preserving video generation (Meng et al., 2026b). The completeness of those observations is therefore an appropriate design question, but the present analysis does not test its reference memory or generated clips. It examines connectivity among the positive pairs in the public LFW assessment list. The contribution is a precise description of which image relationships are explicit in that file. It supplies neither a new identity inference nor permission to augment a restricted training protocol with unlisted pairs.

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.

## 2. A graph built only from explicit positive pairs

Each distinct image key is a node, constructed from its person label and image index. Each of the 3,000 matched rows adds an undirected edge between its two photographs. Connected components are then calculated from these explicit edges. Negative pairs are not used to join components, because their role is a different identity comparison. Separately, the full set of 6,000 rows is checked for duplicate unordered image pairs. There are no duplicate pair rows under that definition.

The positive graph contains 4,720 images and 1,878 connected components. Component sizes are two for 1,259 components, three for 386, four for 159, five for 46, six for 21, seven for five, eight for one, and nine for one. The component count exceeds the 1,532 distinct positive identity labels. Thus a single

person label can be represented through more than one disconnected set of explicitly paired images. A connected component is not a count of people, and a person label does not establish that all of that person's photographs are connected in the observed pair graph.

A large collection of measurements does not guarantee a large collection of independent cases. Repeated observations can share unrecorded acquisition conditions or a common source, making row-level uncertainty misleading. White's covariance estimator addresses heteroskedasticity within a specified regression (White, 1980), while cluster methods address a declared form of grouped dependence (Cameron and Miller, 2015). Neither approach repairs a poorly defined outcome or establishes representativeness of the observed sample. A report should identify the number of groups, their relative sizes, and the reason for selecting the grouping level. When only a few independent groups are available, apparently precise numerical intervals warrant particular caution. This is a limitation of the evidence structure, rather than a reason to count correlated measurements as fresh observations.

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.

A validation split defines which form of transfer has been observed. Related observations can share subjects, scenes, prompts, or acquisition conditions, so randomly splitting rows may leave important information on both sides of the boundary. Structured cross-validation addresses these dependencies (Roberts et al., 2017). Model selection introduces another boundary: choosing a representation, threshold, or tuning parameter using the final assessment data makes the resulting error estimate optimistic (Varma and Simon, 2006). The appropriate held-out unit follows the intended future use. Generalization to a new frame, a new trajectory, and a new subject are not equivalent claims. A study should record those units before fitting or choosing a scoring rule, and preserve an assessment set that has not become part of iterative development.

## 3. What connectivity means for a reference collection

The graph describes comparison organization rather than the visual difficulty of any edge. A component with several images may include nearly redundant views or substantial variation; the metadata alone does not identify which. A disconnected set within one label does not indicate an annotation error or an algorithmic failure. It shows that the protocol did not explicitly pair every available relationship. This is a useful distinction to preserve when a reference system claims to represent a subject broadly.

For generated-subject evaluation, a collection can contain many reference photographs while leaving important relationships weakly observed. A proposed audit could record viewpoint, acquisition, and temporal links where annotations are available, then distinguish those observations from an assumed complete identity manifold. It could also identify which reference subsets a model actually receives. The present static graph cannot establish coverage of yaw, expression, or occlusion because those variables were not measured in this analysis.

Transitive reasoning also needs a protocol boundary. A researcher may be able to infer a candidate relationship from two positive edges, yet that does not mean the inferred pair is an independently observed judgment or a permitted addition to a restricted training set. Any new use should be described as a change in data access or task definition. Otherwise, a performance comparison can reward a method for additional label information while attributing the gain only to its architecture. The audit here performs no such training or augmentation.

The same caution applies to generator memory. Connecting references through a representation does not itself demonstrate that every required appearance relation is preserved in output video. A direct test would need generated trajectories and an independent outcome definition. Connectivity can organize that test and identify gaps, but it cannot replace the outcome. Its scientific value lies in describing the evidence available to the conditioning process and clarifying where an evaluation must collect something further.

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.

An association can suggest a research direction without identifying an intervention. Causal diagrams make it possible to state which variables precede the proposed change, which are consequences, and which selection processes remain uncontrolled (Greenland et al., 1999). Their value lies in exposing assumptions rather than decorating a regression with a causal label. A second issue arises when many outcomes or subgroups are explored: selecting the most favorable comparison can exaggerate the impression of support. False-discovery procedures address a defined family of inferential tests (Benjamini and Hochberg, 1995). They do not turn an exploratory pattern into a replicated mechanism. A careful report distinguishes prespecified comparisons, descriptive sensitivities, and hypotheses requiring new observations. This distinction is especially important when an evaluation offers many plausible metrics.

An automated measurement should not inherit the authority of its training labels without a direct assessment of use. The relevant checks include systematic disagreement with the intended procedure (Bland and Altman, 1986) and calibration of any probability claims (Guo et al., 2017). Ranking performance, absolute accuracy, and confidence reliability are distinct outcomes. A study can improve one while leaving the others unchanged. Keeping the measurements separate permits examination of those tradeoffs and avoids implying that one favorable statistic establishes all aspects of validity. When independent judgments are unavailable, the limitation should remain part of the result. The absence of a target observation cannot be repaired by combining several correlated proxies into a more elaborate total score.

## 4. Connecting an identity claim to its observations

Video evaluation should preserve the distinction among identity, temporal behavior, and general visual appearance. VBench offers a multidimensional approach to video assessment (Huang et al., 2023), while structural similarity (Wang et al., 2004), learned perceptual metrics (Zhang et al., 2018), and generative video-distribution evaluation (Unterthiner et al., 2018) provide different kinds of measurements. Their outputs can disagree for scientifically meaningful reasons. A temporal average may dilute a brief subject substitution, and a strong reference match may reward a nearly static output that fails to carry out the requested motion. The test should therefore retain where and when each observation was made. Human judgments or task-specific measurements may be needed to determine whether a disagreement is consequential. Combining metrics is reasonable only when the combination's interpretation is explicit. A total score should not conceal the fact that one component was unavailable or inappropriate for the very cases motivating the new generation method.

A useful reliability claim contains both a performance statement and a scope statement. Selective prediction shows why a system that answers only some cases must report how many it accepts (Geifman and El-Yaniv, 2017). Confidence estimation under a changed data distribution requires independent checking (Ovadia et al., 2019). These requirements apply to generators and to the automated judges used to assess them. If a judge cannot evaluate a profile, an occluded object, or an unusual interaction, that inability should remain visible rather than removing the case from the generator's score silently. Recording the entire decision path permits an analyst to identify whether an apparent improvement came from better outputs, narrower eligibility, or a more restrictive measurement procedure.

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

LFW's explicit positive-pair graph has more connected components than represented positive identities, while the complete pair list contains no repeated unordered pair. These observations clarify the difference between labels, photographs, and directly listed comparisons. Reference systems for subject-preserving generation can use a similar audit to document their inputs without overstating what connectivity establishes. A graph of available evidence is a useful design object; measured fidelity in generated video remains a separate empirical requirement.

## 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.