Reference Diversity and Redundancy in Identity Mosaics

Blake Peterson (NP), Nathaniel Harrison (BY), Jacob Reynolds (MT)

Abstract

More reference tiles do not necessarily provide more independent identity evidence. This conceptual analysis distinguishes the size of an identity mosaic from its coverage of views, expressions, and acquisition conditions. It proposes matched-budget comparisons that retain redundant and complementary reference sets as separate conditions. ARGUS motivates the structured-reference mechanism, but no mosaic is constructed or tested here. The argument is that an evaluation should connect a claimed multi-view advantage to the information added by the references, rather than attribute every improvement from a larger or better-selected input set to its arrangement alone.

1. A collection can be large and narrow

Several references can repeat essentially the same view while differing in small visual details. Another collection of the same size can expose different sides, expressions, or visibility conditions. Both contain the same number of images, but they supply different evidence for a requested output. An evaluation of reference organization should retain that distinction. Otherwise, a method may appear to benefit from its arrangement when it actually receives more useful input information.

ARGUS organizes selected identity observations into a stacked multi-view conditioning representation (Meng et al., 2026b). This motivates examining redundancy and complementarity in the inputs. The present article does not claim a measured relationship between mosaic diversity and its published performance. It proposes a research design through which that relationship could be tested. No references are selected, tiled, embedded, or used to generate video here. The objective is to make the informational comparison explicit before an architectural explanation is assigned.

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.

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. Matching the budget without assuming equal evidence

A proposed experiment would fix the number and nominal resolution of reference observations while varying a declared coverage property. One condition might repeat similar views; another might include complementary views. The source and quality of the observations should be documented, because a coverage manipulation can also change lighting, blur, or expression. An independent annotation or acquisition record would be needed to establish which property actually differs. A label such as diverse should not substitute for that evidence.

The experiment should distinguish reference selection from reference organization. A common candidate pool could be selected by different procedures and then arranged under a fixed mechanism. Separately, the same selected observations could be presented through alternative organizations. These comparisons answer different questions. If both operations change together, a favorable output cannot isolate which one mattered. The study may still evaluate a complete workflow, but its explanation should remain at that workflow level.

Input budget also includes information outside the image tiles. Prompt text, a first frame, manual annotations, and adaptation examples can influence how useful a reference set becomes. A comparison should record those inputs and any manual curation. A system that receives a carefully chosen target view may face an easier task than one supplied with an uncurated pool. Equal image counts alone would not make their access equivalent. The evaluation should preserve that limitation instead of hiding it behind a common nominal mosaic size.

The recognition literature provides both scoring methods and warnings about incomplete evaluation populations. FaceNet's learned embedding (Schroff et al., 2015), ArcFace's angular-margin formulation (Deng et al., 2018), and CosFace's margin-based discrimination (Wang et al., 2018) are alternative measurement foundations rather than interchangeable proof of personal identity. VGGFace2 introduces substantial pose and age variation (Cao et al., 2017), MegaFace studies recognition at large scale (Kemelmacher-Shlizerman et al., 2015), and FairFace emphasizes representation in annotated face datasets (Kärkkäinen and Joo, 2019). An identity-preserving generator should be assessed with the relevant dimensions recorded separately. A result from familiar frontal images cannot establish reliability under a missing viewpoint. Nor can a frequency-balanced set establish demographic balance without the required labels. The evaluation should retain failures of detection or measurement, identify what threshold was used, and explain whether the endpoint concerns verification, identification among distractors, or human recognizability over a moving sequence.

A proposed explanation should be distinguishable from alternative explanations using observations the study can actually collect. Diagramming the relationships among selection, exposure, measurement, and outcome makes that requirement explicit (Greenland et al., 1999). The diagram does not establish its own correctness, but it prevents adjustment choices from remaining invisible. The same transparency is needed when investigators inspect many endpoints before choosing a narrative. Controlling false discovery concerns the behavior of a specified testing family (Benjamini and Hochberg, 1995); it is not a general certificate of scientific validity. Primary endpoints, secondary descriptions, and future mechanism tests should be named separately. Readers can then assess whether an attractive pattern was predicted, explored, or merely noticed after the available comparisons were examined.

The phrase generalization needs an object: another observation of the same source, another source within the same collection, or a different acquisition process. Each requires a different assessment boundary. Structured validation makes explicit how correlated observations are assigned across training and evaluation (Roberts et al., 2017). Selection bias can still arise if the final boundary is repeatedly consulted

while choosing a method (Varma and Simon, 2006). Publishing the split identifiers and the information permitted during each stage allows the claimed transfer to be checked. When a study proposes a new architecture but performs only a data audit, it should state that difference directly. Useful evidence about data coverage does not require pretending that a prediction system has already been trained and externally validated.

3. Connecting coverage to a requested output

The value of a reference depends partly on the intended generation. A profile observation may be especially informative for a profile request, while another task may depend more on an expression or transient appearance detail. The study should therefore link reference coverage to requested conditions and report performance within those conditions. A single average across convenient prompts may conceal whether the added evidence helps where the proposed mechanism is supposed to matter.

Redundancy is not automatically useless. Repeated observations can reduce ambiguity about a noisy view or provide variation in acquisition quality. A design should therefore avoid treating duplicate-looking inputs as a universally inferior control. It should state the question being tested and preserve measurement uncertainty. The outcome could reveal a tradeoff between breadth and confidence in a smaller set of observations. That tradeoff requires actual outputs and independent judgments; it is not established by the number of distinct reference filenames.

The report should retain failed or unassessable cases and identify whether reference coverage changes assessment availability. A method could improve apparent identity scores by producing only familiar views even when unfamiliar ones were requested. Separate viewpoint compliance and identity outcomes would expose that behavior. The most useful result would connect a declared input difference to its relevant output consequence while keeping other access and operating conditions recoverable.

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.

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. Reference access and measurement reliability

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.

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.

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

An identity mosaic should be described through both its budget and the evidence its observations cover. The proposed design separates selection, organization, and task-specific usefulness, while recognizing that redundancy can sometimes provide legitimate information. It reports no new result for ARGUS. Its contribution is a way to test whether a multi-view advantage follows from a declared conditioning mechanism, improved reference access, or a complete workflow that combines both.

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.