Diversity among Generations of the Same Referenced Subject

Henry Parker (GM), Leo Pearson (MN), Dennis Rogers (CY)

Abstract

Producing several distinct videos of one person is useful only if their variation remains compatible with the intended identity and task. This article proposes studying diversity conditional on subject preservation and instruction execution. ARGUS motivates the reference-conditioned setting. No samples are generated and no diversity statistic is calculated. The design keeps variation in pose, motion, scene, and incidental appearance separate, preventing either near-duplicate outputs or unconstrained changes of subject from receiving an unexplained favorable diversity label.

1. Variation needs a valid region

A generator can produce different pixels by changing a background, moving the camera, varying a facial expression, or replacing the person. These sources of variation have different scientific meanings. A generic distance among outputs cannot decide which changes are desirable. Before measuring diversity, an evaluation should state the task's stable requirements and the dimensions in which variation is welcome. Otherwise a failure of identity preservation may appear as a successful increase in output diversity.

ARGUS organizes multi-view references to preserve a subject in generated video (Meng et al., 2026b). Its setting motivates asking what variation remains possible once the subject and instruction are held fixed. This article outlines a conceptual sampling study rather than evaluating ARGUS generations. It does not calculate a diversity score or claim that a conditioning mechanism limits variation. The aim is to connect a future diversity result to the particular freedom that the generation task intends to allow.

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. Sampling outputs without selecting the conclusion

A proposed comparison would specify subjects, instructions, reference sets, model versions, sampling settings, and the number of attempts before generation. Every attempt should receive a record, including failures and outputs that cannot be assessed. If a user-facing workflow selects among outputs, its selection rule and effort should be evaluated explicitly. Reporting only the final selected clips would otherwise leave the relationship between sample diversity and search opportunity impossible to reconstruct.

The outcome table would include subject recognizability and instruction execution before interpreting variation among successful samples. Conditional diversity is useful, but its conditioning should be visible: a method that produces very few valid outputs can appear diverse within that small retained subset. The report should therefore present validity rate and conditional diversity together. It should also show how uncertain identity decisions affect the retained set rather than treating all excluded observations as irrelevant to the comparison.

Different variation dimensions need corresponding measurements. Motion diversity is not established by a change in color, and scene diversity is not established by numerical noise in otherwise identical clips. Independent labels or task-linked features may help, provided they are validated for the generated population. A small set of interpretable dimensions can be preferable to a single large feature-space distance whose sensitivity to the intended changes has not been examined.

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.

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.

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.

3. Comparing diversity with preservation and effort

The number of sampled outputs affects many diversity summaries. Comparing a large candidate pool with a small one can reward sampling effort rather than a generation mechanism. A study should use matched sample counts or present diversity as a function of declared effort. Repeated draws from the same subject and prompt should remain grouped for uncertainty calculations. Many pairwise distances among a few outputs do not create a correspondingly large set of independent generation observations.

The relationship between preservation and diversity can be reported as a profile rather than resolved through an arbitrary total. A method may preserve the subject closely while varying little, whereas another may produce richer valid motion with a somewhat different visual resemblance. The application determines which tradeoff matters. Any combined utility should make that preference explicit and retain the component outcomes so that another user can interpret the same observations under a different requirement.

Finally, variation across reference sets should remain distinct from variation across random samples. Changing the supplied views changes the available evidence and may change the identity interpretation itself. A proposed study could cross these sources of variation, but it should not pool them indiscriminately. Separating reference choice, instruction choice, and generation randomness makes it possible to identify where diversity enters the workflow and which part of the resulting range a user can actually control.

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.

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

Subject preservation is one dimension of a larger video-generation task. VBench explicitly motivates separating evaluation dimensions (Huang et al., 2023), and the visual literature offers structural (Wang et al., 2004), perceptual (Zhang et al., 2018), and temporal-distribution measurements (Unterthiner et al., 2018). A study should choose among them according to the claim and available reference information. When identity, motion, and visual quality move in different directions, the disagreement should be analyzed rather than automatically resolved by a weighted total. The weighting itself expresses a preference that may differ across applications. Preserving per-frame and per-clip records allows alternative interpretations to be evaluated later. It also makes coverage visible when some poses or occlusions prevent valid scoring. A generator can demonstrate a useful, limited preservation capability without having established universal reliability across all subjects, durations, and visual conditions.

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.

An evaluation remains comparable only if its inputs and transformations can be identified. The FAIR principles emphasize persistent identification and reusable scientific metadata (Wilkinson et al., 2016). Work on machine-learning technical debt highlights how dependencies and hidden couplings accumulate in a pipeline (Sculley et al., 2015). In this setting, a score should be linked to the exact references, generation settings, evaluator version, and aggregation rule that produced it. A replacement checkpoint or a revised eligibility filter may change the scientific question even when the output column keeps the same name. Preserving such changes in a release record permits later comparisons to be interpreted honestly. The objective is not maximum documentation volume, but a recoverable account of the decisions that affect the result.

5. Conclusion

Diversity is useful evidence when its permitted dimensions, validity conditions, sample counts, and effort are declared. The proposed protocol studies variation among referenced-subject outputs while retaining preservation and instruction outcomes. It presents no new diversity estimates. Its contribution is a framework in which distinct valid performances can be recognized without rewarding subject substitution, hidden selection, or a larger sampling budget as though each were the same improvement in generative range.

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.