Viewpoint Coverage before Large-Yaw Identity Evaluation

Noah Hoffman (TL), Caleb Barrett (QA), Lucas Evans (GH)

Abstract

A large-viewpoint test is difficult to interpret when the reference set's own viewpoints are unknown. This conceptual analysis proposes recording reference and output orientation separately, together with the uncertainty of any viewpoint annotation. ARGUS motivates the problem through its emphasis on multi-view identity evidence and viewpoint stress. No pose estimator or generation model is run here. The proposed study distinguishes missing reference support, failed generation, and failed identity measurement, so a profile-view score can be read as evidence about a declared comparison rather than an unrestricted preservation capability.

1. A difficult output may be familiar in the reference

A generated profile can be an interpolation from a supplied profile reference or a more demanding continuation from frontal observations alone. Both are legitimate generation tasks, but they provide different evidence about the conditioning mechanism. Calling both a large-yaw test does not make their information access equivalent. The evaluation should therefore describe the relationship between requested viewpoint, supplied reference views, and the output actually obtained.

ARGUS organizes multi-view identity information and includes viewpoint-oriented stress in its evaluation motivation (Meng et al., 2026b). The present article does not reinterpret its reported scores or infer unobserved reference distributions. It develops a research design for making viewpoint coverage explicit. No facial orientation, recognition score, or generated clip is calculated here. The goal is a recoverable specification that can distinguish a useful limited result from a broader claim that the available views do not support.

A reference collection can grow in several ways: more identities, more images of existing identities, or more comparisons among the same images. The LFW protocol makes these units distinguishable (Huang et al., 2008), while its official loader documentation preserves the difference among development and assessment files (Pedregosa et al., 2011). Datasheets provide a framework for explaining why those objects were collected and which uses they support (Gebru et al., 2021). A subject-generation study should apply the same discipline to reference mosaics and output trajectories. The number of reference tiles is not automatically the number of independent observations, and the number of scored frames is not automatically the number of subjects. Recording the hierarchy permits appropriate weighting and uncertainty calculations later. Where a study uses only a public pair list, its conclusions should remain about that list and its protocol. They cannot supply observations of appearance, motion, or recognition errors that were never processed.

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.

2. The annotations that define the viewing condition

A proposed record would identify each reference observation and the viewpoint information available for it. That information might come from a calibrated acquisition, an independently supplied annotation, or a declared estimator. These are different measurement sources, and their uncertainty should remain visible. An estimated angle is not automatically exact ground truth. If difficult views also defeat the orientation estimator, the assessment should retain that loss of coverage instead of treating the missing values as ordinary frontal cases.

The output record should identify the requested orientation and the realized viewing condition. A generator that avoids the requested profile may retain identity simply by remaining near a familiar reference view. Scoring only the realized easy view would reward failure to execute the instruction. The evaluation therefore needs separate endpoints for viewpoint compliance and identity preservation, with a rule for outputs in which one or both cannot be measured. Their combination should be interpretable before the preferred model's clips are inspected.

Reference diversity should also distinguish number from coverage. Several nearly identical photographs do not provide the same information as a small collection of genuinely different views. A useful comparison would preserve the number, quality, and distribution of reference observations, then vary the relevant viewpoint support through a declared design. The experiment should avoid claiming an isolated effect of multi-view organization when the proposed method also receives additional images or a better-selected reference.

Replication has several levels: a second measurement, a second trajectory, and a second independently acquired setting are different additions to evidence. Statistical calculations should retain those distinctions. Heteroskedasticity-robust regression inference concerns unequal conditional variances (White, 1980); cluster-robust inference concerns dependencies shared within groups (Cameron and Miller, 2015). The group definition should be justified through the data collection and evaluation process, and not chosen because it produces a preferred significance result. Unequal group sizes also influence what a pooled average represents. Reporting both the group count and observation count permits a reader to see whether the calculation describes common cases, heavily recorded cases, or an explicitly balanced group population. No variance estimator alone resolves that choice of estimand.

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.

A subject-generation pipeline can succeed at using a reference without succeeding at every identity requirement. DreamBooth's model adaptation (Ruiz et al., 2022) and IP-Adapter's image prompting (Ye et al., 2023) offer examples of reference use with different costs and information access. CLIP (Radford et al., 2021) and DINO (Caron et al., 2021) provide further visual representations whose similarity scores

may emphasize other aspects of content. The evaluation should therefore define its target before choosing a convenient feature. If the intended target is a recognizable moving person, that target needs appropriate observations across the relevant views and expressions. A high generic image similarity can instead reflect scene layout or texture. Matching reference budgets and documenting the evaluator's representation permit those explanations to be compared. The objective is to connect a claimed conditioning advantage to the particular identity evidence the resulting output actually preserves.

3. A preservation result conditional on assessability

An identity instrument can have different behavior on frontal and profile observations. A decrease in its score can reflect the generator, the measurement procedure, or both. A fixed set of outputs assessed by independent procedures would help distinguish those explanations. Human observers should receive appropriate reference evidence and task instructions, and they should be allowed to express uncertainty where recognition is genuinely ambiguous. The study should not demand a confident verdict merely to obtain a complete table.

The relevant denominator includes scheduled cases that fail generation or assessment. A method can appear strong if it produces valid profiles only in easy settings while difficult outputs become unscorable. Reporting coverage by requested viewpoint would expose that pattern. It would also permit an application to identify a useful operating region without asserting universal invariance. A limited result is scientifically meaningful when its conditions and missing evidence are stated precisely.

The test should include varied subjects and scene conditions, while protecting a boundary between development and final assessment. Choosing reference views after examining the final failures would change the information available to the model. Such diagnostic improvement can be useful, but its evaluation should use new cases. The record should preserve when references were selected and which observations guided that selection, so the eventual viewpoint claim remains tied to an actual information boundary.

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.

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

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.

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

A viewpoint-preservation result should identify what the model saw, what it was asked to generate, what it produced, and what the evaluator could validly judge. The proposed design separates reference support, viewpoint compliance, identity, and assessment coverage. It supplies no new result for ARGUS. Its contribution is to make a large-yaw test interpretable as a bounded experiment rather than allow a difficult-sounding label to substitute for a documented comparison.

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.