# Designing Human Comparisons for Subject Consistency

Austin Warren (HR), Simon Collins (LA), Jeffrey Mason (LI)

## Abstract

Human assessment of subject consistency is sensitive to the references, questions, and viewing conditions presented to observers. A preference for an attractive clip may not answer whether it preserves a particular person. This conceptual article develops a comparative study in which identity, instruction execution, and general appearance receive separate judgments. It uses ARGUS to motivate evaluation with multiple reference views. No participants are recruited and no preference rates are reported. The proposed design makes a human comparison answer a specific preservation question.

## 1. Asking observers a question they can answer

The phrase same subject can refer to different perceptual requirements. An observer might recognize a face despite changed clothing, or might identify a character largely through a costume and background. Neither interpretation is automatically wrong, but they correspond to different tasks. Before collecting judgments, a study should describe which evidence is supposed to persist and which transformations the generator is expected to carry out. Otherwise its preference score can conceal disagreement about the task itself.

ARGUS presents a reference-conditioned video setting in which organized views contribute identity evidence (Meng et al., 2026b). A human comparison in this setting should state which of those references the observers see. The present article proposes such a comparison without assessing any generated clips. Its concern is whether the viewing protocol permits a subject-consistency interpretation, especially when a method produces appealing motion that can distract from a subtle change in identity.

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.

The validity of an identity evaluation depends partly on how its scientific objects are named and linked. LFW's identity, image, pair, and fold units are explicit parts of its protocol (Huang et al., 2008). The scikit-learn implementation documents which file corresponds to each supported task (Pedregosa et al., 2011). Dataset provenance should similarly make the intended use and selection process recoverable (Gebru et al., 2021). In video generation, an evaluator needs additional links among reference observations, subjects, clips, and frames. Repeated references can provide useful evidence without becoming independent subjects. A pair-level metadata audit can reveal those relationships, but it cannot evaluate the visual content of a generated clip. The distinction is productive: data organization can be checked before an expensive model study begins, and the missing measurements can be specified rather than silently assumed to exist.

## 2. Structuring the comparison and its records

A proposed task would show the relevant references together with clips generated for a common instruction. Observers would first judge recognizability of the referenced subject under a defined question. Separate questions would assess the requested transformation and general visual coherence. The response format should allow no discernible difference and insufficient evidence. Forcing a binary winner in ambiguous cases may create apparent information that the observer did not actually possess.

Presentation order, clip position, playback opportunity, and visible identifying labels should be controlled and recorded. Blinding method names can reduce an obvious source of expectation, although it does not prevent observers from recognizing characteristic artifacts. A pilot could test whether instructions are understood and whether the viewing duration is adequate. Its purpose would be to improve the procedure; pilot observations should not be silently combined with a later confirmatory comparison after the questions have changed.

Sampling should preserve the hierarchy of subjects, instructions, generated samples, and reviewers. Several reviewers of one clip add information about judgment variability, not additional independently generated clips. Several clips of one subject likewise do not create new subjects. Assignment of review tasks can distribute fatigue and balance comparisons. The final analysis should retain these links so that uncertainty reflects the actual evidence rather than treating every button press as an unrelated experimental unit.

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.

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.

## 3. Reading agreement and preference together

Agreement among observers can help describe whether the task has a stable interpretation, but unanimous judgments do not establish validity by themselves. Observers can share a cue that the study did not intend to measure. Follow-up questions or carefully chosen contrasting cases may reveal whether they rely on face structure, clothing, motion, or scene context. Such checks should be planned and should avoid inferring a general cognitive mechanism from a small convenience sample.

A paired preference result should be accompanied by the conditions on which it is based. A method could win often on familiar frontal references yet show no advantage during occlusion. Another could preserve the person while failing the requested action. Reporting these outcomes separately allows the reader to judge the tradeoff. A single weighted total is possible, but the weighting expresses a use preference and should not be represented as a neutral definition of identity.

Reviewer recruitment and exclusions also affect interpretation. A study should document eligibility, relevant experience, training, incomplete tasks, and exclusion rules without exposing unnecessary personal information. Results from one panel support conclusions about that panel and its task conditions. Transport to another population of users requires additional evidence. This is particularly important when recognizing the subject depends on prior familiarity, cultural conventions, or specialized knowledge of a character rather than solely on the supplied photographs.

Different reference-conditioning mechanisms can receive different amounts of information. DreamBooth adapts a generation model around a subject (Ruiz et al., 2022), while IP-Adapter offers a compatible image-conditioning route (Ye et al., 2023). A fair comparison should preserve how many views, which resolutions, and which adaptation opportunities are supplied. CLIP's language-supervised visual representation (Radford et al., 2021) and DINO's self-supervised features (Caron et al., 2021) can provide useful descriptions of outputs, but their training objectives do not establish a person-recognition standard. The endpoint should specify whether success means matching reference appearance, retaining a recognizable subject, or satisfying a requested scene. Those properties can disagree. A useful analysis would keep the disagreement visible and ask which source of evidence each method uses. The relevant advantage is a measured improvement under a declared information budget, not an unqualified increase in similarity to whatever reference happened to be available.

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.

Coverage should be reported whenever the system or its evaluator can avoid difficult cases. An abstention policy changes the population on which an accuracy or fidelity score is calculated. Selective classification formalizes the joint reporting of coverage and conditional performance (Geifman and El-Yaniv, 2017). Distribution-shift research adds that estimated uncertainty can behave differently outside the development population (Ovadia et al., 2019). A benchmark should therefore retain which observations were eligible, accepted, rejected, or unavailable. A result on a carefully limited subset can be useful, provided the subset

is visible. The challenge is to prevent a narrow success claim from becoming a broad capability claim through omission of the denominator or the circumstances in which the system declines to operate.

## 4. Designing an interpretable preservation study

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 substitute measurement needs evidence about its relationship to the intended judgment. Two scores may correlate strongly while differing systematically in particular conditions. Method-comparison analysis distinguishes association from agreement (Bland and Altman, 1986). When a system additionally reports confidence, calibration asks whether those probabilities correspond to observed outcome frequencies (Guo et al., 2017). These are separate properties: a metric can rank examples usefully while having a distorted numerical scale or unreliable confidence. An evaluation should therefore retain the original target observations when possible, describe the scale of each endpoint, and examine disagreements instead of averaging them away. A convenient automated score becomes more credible when its limitations are measured in the conditions that motivate its use.

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

A useful human comparison begins with a defined identity question and an adequate reference presentation. It preserves ambiguity, separates appearance preferences from subject consistency, and accounts for repeated observations of clips and reviewers. The design proposed here has not been implemented. Its value is to make future preference evidence interpretable: a winning clip should win on the stated preservation requirement rather than on an unspecified mixture of attractiveness, motion, familiarity, and presentation effects.

## 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.