Agreement between Human and Automated Subject Judgments

Frank Crawford (IT), Nicholas Duncan (ZM), Adam Campbell (OM)

Abstract

An automated identity score becomes more useful when its relationship to human recognizability is examined under the intended video conditions. Agreement alone, however, does not show that both assessors answer the same question. This article proposes a comparison with shared reference records, independent judgments, and retained uncertainty. ARGUS motivates the subject-preserving setting. No human responses or metric outputs are collected. The design separates measurement agreement, operating decisions, and the task consequences that make a disagreement scientifically meaningful.

1. Agreement requires a common question

A human observer can judge whether a person is recognizable, whether an output resembles a reference, or whether one clip is preferable overall. An embedding score may correspond more closely to one of these questions than another. Comparing their outputs without specifying the task can produce a numerical agreement measure whose meaning remains unclear. A validation study should first identify the intended decision and the information available to both kinds of assessor.

ARGUS uses organized identity references to guide video generation (Meng et al., 2026b). That setting motivates examining whether automated and human judgments support a common preservation interpretation. The present article proposes a validation study and does not assess ARGUS clips. It reports no correlation, agreement rate, or calibrated threshold. Its objective is to make those future quantities interpretable by linking them to declared references, viewing conditions, and a separately defined judgment of the intended subject.

Precision and identification are separate properties of a result. A narrow interval can surround an association whose scientific target is unclear. Robust covariance calculations allow particular departures from simple error assumptions, including unequal variance (White, 1980) and dependence within declared clusters (Cameron and Miller, 2015). Their role is to improve uncertainty calculations under those assumptions. They do not establish a causal assignment, guarantee a correct measurement, or expand coverage to conditions missing from the data. The analyst should therefore explain both the averaging population and the source of replication. An evaluation repeated many times on one history may provide information about stochastic variability while supplying little evidence about transfer to another history. Those two sources of uncertainty should remain distinguishable in the final interpretation.

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. Collecting comparable but independent observations

A proposed dataset would retain generated clips, source references, instructions, model versions, and sampling records. Human reviewers would receive a declared presentation and answer a specific recognizability question, with uncertainty allowed. Their responses should be collected independently of the metric output. Showing the automated result before judgment would change the task and could manufacture agreement through influence rather than establish that the instrument independently captures a useful aspect of the video.

The automated path would preserve its detector, representation, preprocessing, threshold, and aggregation rule. An unavailable score would remain a recorded status. If humans receive references unavailable to the metric, the difference should be explicit. The study could deliberately examine several information conditions, but their results should remain separate. A mismatch in evidence access is not merely random error; it can change what each assessor is capable of deciding about the same output.

Development and final assessment should be separated when thresholds or mappings are selected. Fitting a relationship and reporting its apparent agreement on the same examples would not demonstrate transfer to new observations. Splits should match the intended generalization unit, such as subjects or tasks. Repeated frames and multiple reviewers should retain their links. A large number of paired responses from a small number of clips does not establish the same evidence as a broad sample of independently generated subjects.

Describing a contrast accurately can be more valuable than giving it an unjustified mechanistic name. A causal interpretation needs assumptions about how exposure and outcome arose, including selection into the analyzed sample. Causal diagrams provide a language for those assumptions (Greenland et al., 1999). If several alternative metrics or partitions are examined, the investigator should also account for the freedom to select whichever result appears strongest. False-discovery control is relevant to explicit collections of statistical tests (Benjamini and Hochberg, 1995), while a descriptive sensitivity analysis should be presented as such. A model can fit the observed rows well and still leave the principal explanation unresolved. Reporting that boundary preserves the value of the observation without overstating what adjustment or a significance threshold has established.

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.

3. Understanding disagreement rather than hiding it

The analysis would retain response distributions and examine disagreements by documented condition. A high overall correlation can coexist with consequential errors near a decision threshold. Conversely, different numerical scales can yield similar decisions under appropriate operating points. Reporting both continuous relationships and declared decisions can clarify that distinction. The study should avoid choosing a favorable statistic after observing which summary makes the automated instrument appear closest to the human panel.

Human judgments have their own limitations. Reviewers can disagree, rely on incidental cues, or lack sufficient evidence in difficult views. A majority label is useful for some purposes but should not erase those observations. Adjudication can be documented as another stage rather than treated as proof that every initial uncertainty has disappeared. When both human and automated assessment are weak, the appropriate result may be limited coverage instead of a forced conclusion about which side is correct.

The final claim should connect agreement to the intended use. A metric suitable for screening obvious substitutions may not be suitable for ranking subtle identity differences or deciding every individual case. The validation population also matters: results on one generator's photographic outputs may not transfer to stylized releases or other visibility conditions. A bounded validation can still be valuable when it identifies the operating task, the unresolved region, and the observations that would be needed before a broader deployment claim is justified.

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.

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.

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.

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.

Replacing an expensive judgment with an automated score changes the measurement process. A positive correlation is evidence of co-movement, but it does not establish that the two procedures can be used interchangeably (Bland and Altman, 1986). A calibrated forecast has a different requirement: its stated probabilities should match relevant observed frequencies (Guo et al., 2017). Neither property follows merely from a favorable overall ranking. A useful comparison preserves paired measurements, identifies the conditions in which disagreement matters, and keeps scale transformations explicit. If the intended judgment concerns recognizability, playability, or physical compatibility, the proxy should be assessed against that judgment directly. Agreement on easy cases cannot supply evidence about the difficult conditions absent from the comparison.

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

Human-automated agreement should be studied under a shared, explicit preservation question with independent observations and recorded information access. The proposed design reports no new validation result. It preserves thresholds, uncertainty, coverage, and disagreement patterns so that future evidence can establish a useful measurement role without treating a correlation or majority label as universal proof that an automated identity score captures every consequential aspect of subject continuity.

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.