# Clothing Cues and the Measurement of Personal Identity

Patrick Bryant (BA), Edward Hill (KW), Trevor Morgan (FJ)

## Abstract

A familiar outfit can make an output look reference-consistent even when the intended person is not preserved. This conceptual analysis proposes crossing subject and clothing conditions so that identity and wardrobe compliance can be assessed separately. ARGUS motivates the reference-based video setting. No clothing labels, images, or generated clips are analyzed here. The design treats clothing as legitimate task information whose role must be specified, rather than assuming it is always irrelevant or always sufficient. It also retains cases in which visible evidence cannot support a confident identity judgment.

## 1. An informative cue is not a complete definition

Clothing can help an observer follow a person through a scene, especially when the face is small or temporarily hidden. It can also become a misleading cue when the task requires the same person in a different outfit or different people wearing similar garments. An evaluation should therefore identify the role clothing plays in its claim. A reference match based mainly on wardrobe may be useful for appearance continuity without establishing the intended identity requirement.

ARGUS conditions video generation on selected identity observations (Meng et al., 2026b). Those observations can contain clothing and other context in addition to facial evidence. The present article does not claim that the framework relies improperly on such cues. It proposes a diagnostic design separating subject preservation from wardrobe execution. No new image annotation or video experiment is performed. The goal is to make explicit which information an output must preserve and which information the task allows or requires it to change.

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.

## 2. Crossing the reference and requested conditions

A proposed study would document the subject references and the clothing observations present in them. It could then compare tasks that retain an outfit, change it, or require visually similar clothing across different referenced subjects. Each condition needs an independently understood identity and wardrobe target. A change request should be sufficiently specified to assess, but not so overconstrained that the evaluator mistakes permitted stylistic variation for failure. The task should be established before scoring the preferred system.

The endpoints would remain separate. Identity asks whether the output is compatible with the intended subject under the available evidence. Wardrobe compliance asks whether the requested clothing condition is realized. General visual quality and temporal continuity concern further properties. A system can do

well on any subset of these. A high identity-like visual score should not automatically excuse refusal to change clothing, and successful wardrobe editing should not be treated as evidence that the person remained the same.

Reference information must be matched carefully. If a generator receives a photograph of the desired outfit on the intended person, its task differs from generating that combination without such an observation. Manual selection or additional reference descriptions can also change the available evidence. The report should preserve those inputs and any adaptation steps. A complete workflow comparison is legitimate, but its explanation should reflect the information each workflow was given rather than attribute all differences to one conditioning component.

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 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.

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. Assessing identity when wardrobe dominates the image

Some outputs may contain too little facial or other subject evidence for a reliable identity judgment. A consistent outfit can make the scene easy to track while leaving personal identity unresolved. The evaluator should be allowed to record that distinction. Treating a clothing match as a confident identity verdict overstates the observation, while dropping the case silently can inflate assessed preservation. Coverage and the information used for judgment belong in the result.

A fixed output could also be assessed under different permitted evidence presentations. For example, one procedure might judge the whole scene and another use a declared subject region. Such a comparison would examine measurement sensitivity, not necessarily generator behavior. Any crop or masking operation could itself remove useful evidence or introduce artifacts, so its consequences would need independent checking. The design should avoid declaring one procedure universally correct before establishing which task it supports.

Temporal instructions add another requirement. A subject may change an outfit at a particular event, or the clothing may be required to remain stable throughout a scene. A frame-average score can conceal a wardrobe change at the wrong time or a brief subject substitution. The protocol should preserve the requested event order and time-indexed observations. These are proposals for a direct study, not outcomes inferred from the architectural description or from a few familiar-looking reference images.

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.

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 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.

## 4. Connecting an identity claim to its observations

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 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.

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

Clothing is useful visual information whose task role should be declared. The proposed design crosses subject and wardrobe conditions, measures their outcomes separately, and preserves uncertainty when the available image evidence cannot establish identity. It offers no new ARGUS result. Its contribution is to prevent a legitimate appearance cue from silently becoming the entire definition of the person a generation system is supposed to preserve.

## 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.