Journal of Measurement and Evidence
Volume 1 · Number 1 · 2026

One Nearest Reference, Two Claims: Separating Identity from Capture-Pose Agreement

Arthur Crawford[1] · Vincent Collins[2] · Gavin Bell[3]

[1] Ireland · [2] Finland · [3] Sweden

Original publication: October 2, 2026 | Article 697

Abstract

A correct identity match need not be a matching view. We scored the same nearest-gallery retrievals for 155 CMU queries using identity and source head-pose labels. Of the retrieved references, 142 matched both attributes, nine matched identity alone, and four matched pose alone. The equal-person identity-minus-pose agreement difference was 3.13 percentage points, with a working interval spanning zero. Identity-correct but pose-different retrievals averaged 5.63%. The practical distinction survives the uncertain aggregate gap: identity preservation and capture-pose fidelity are separate claims, and their meaning depends on what the comparison is intended to reward.

Construct audit of a fixed image-retrieval result. Identity and pose are source labels attached to photographs; no generated image or learned pose estimator is evaluated.

One Nearest Reference, Two Claims: Separating Identity from Capture-Pose Agreement

> Construct audit of a fixed image-retrieval result. Identity and pose are source labels attached to photographs; no generated image or learned pose estimator is evaluated.

Abstract

A correct identity match need not be a matching view. We scored the same nearest-gallery retrievals for 155 CMU queries using identity and source head-pose labels. Of the retrieved references, 142 matched both attributes, nine matched identity alone, and four matched pose alone. The equal-person identity-minus-pose agreement difference was 3.13 percentage points, with a working interval spanning zero. Identity-correct but pose-different retrievals averaged 5.63%. The practical distinction survives the uncertain aggregate gap: identity preservation and capture-pose fidelity are separate claims, and their meaning depends on what the comparison is intended to reward.

The ambiguity inside a favorable reference match

A reference-based image comparison often returns one similarity score or one nearest photograph. It is tempting to interpret that result as a general measure of preserved appearance. Yet the reference can resemble the query in several ways: it can depict the same person, share the same head pose, or match scene and framing properties. A favorable result for one attribute need not establish another.

We ask whether successful roster-identity lookup also establishes capture-pose agreement when the identical retrieved reference is scored by both source attributes. There is only one retrieval procedure. We change the claim assessed by its output, rather than fit a second classifier or compare two learned scoring systems. This allows attribute agreement to be separated from variation introduced by different retrieval algorithms.

Meng et al., 2026 study multi-view identity mosaics for subject-preserving video generation. Its multi-view setting makes the distinction especially relevant: preserving a person's identity can be desirable even when the view changes. We do not claim that pose differences are defects in that model or reproduce its outputs. Our experiment clarifies how an assessment should state whether it seeks identity invariance, view agreement, or both.

The measurement framework in Stevens, 1946 motivates naming the attribute behind a number. Carroll et al., 2006 further distinguishes an observed proxy from an underlying construct. Here a nearest-image result is the proxy. The constructs assessed are two discrete source labels, with no assertion that a pixel distance directly measures a continuous amount of identity or pose fidelity.

Two legitimate objectives can value the same result differently

If the objective is recognizing a person across viewpoints, a reference of the correct identity at another pose is a success. If the objective is reproducing a particular view, that same reference provides incomplete evidence. Conversely, a photograph of another person in the correct pose does not satisfy an identity objective even though it agrees on the capture category. No universal scalar ordering follows without specifying which objective matters.

Identity-oriented representations such as Schroff et al., 2015, Deng et al., 2018, and Wang et al., 2018 are designed around identity discrimination rather than exact view reproduction. Their purpose illustrates why

pose variation may be something to tolerate instead of penalize. We do not run those embeddings, and their learned invariances cannot be assumed to characterize the raw-pixel rule used in this experiment.

Quality-aware approaches in Kim et al., 2022 and Meng et al., 2021 introduce another distinction. Image quality can affect recognition, but quality and correct identity are still different assessment targets. A reference can be visually sharp while showing the wrong person, or be difficult to compare while retaining the same identity. Our source labels do not provide an image-quality ground truth for adjudicating those cases.

Dataset design also determines which attributes can be assessed. Cao et al., 2017 incorporates substantial pose and age variation, Huang et al., 2008 supplies a different unconstrained face setting, and Kemelmacher-Shlizerman et al., 2015 emphasizes a large recognition scale. Kärkkäinen and Joo, 2019 addresses balanced attribute coverage for its own objectives. These sources motivate explicit construct and population descriptions, rather than treating a small familiar-roster experiment as a general face-recognition or fairness assessment.

The one retrieval operation used here

The public images come from Mitchell, 1997, with original educational context and naming conventions in Carnegie Mellon University, 2026. Twenty source identities appear under combinations of head pose, expression, and eye condition. We decode quarter-size grayscale images, thirty by thirty-two pixels, from the usable source capture set. Source-marked camera glitches remain excluded; no missing capture is imputed.

The gallery contains 158 open-eye images with neutral or happy expressions, spanning the available poses. The 155 queries have open eyes with sad or angry expressions. All query identities belong to the gallery. Capture overlap is avoided through the expression partition, but this remains a closed-roster task with familiar identities and a shared source collection setting.

Pixels are divided by 255 and compared through squared Euclidean distance over the whole image. The nearest reference supplies the retrieved photograph, with lexicographic gallery ordering resolving an exact tie. No learned embedding, face alignment, image centering, or pose-estimation network is involved. Background, clothing, and framing can therefore influence which reference is closest.

Each retrieved reference has an identity label and a categorical head-pose label parsed from its filename. Identity agreement means that the retrieved and query identity labels are equal. Pose agreement means equality of the source categories straight, left, right, or up. It does not mean equal measured yaw and pitch angles, a three-dimensional reconstruction match, or identical facial landmarks.

The same reference index is used for both scores. The procedure never searches again for a pose-matching reference after the identity result is known, or vice versa. This rule matters because separately optimizing the two retrievals could produce favorable scores that no single selected reference actually satisfies. Our joint event table keeps that relationship visible.

Joint events rather than two detached percentages

For each query, we retain four possible outcomes: both attributes agree, identity alone agrees, pose alone agrees, or neither agrees. These categories exhaust the defined comparison without assigning an unobserved perceptual preference between them. In particular, identity-only is not labeled an error unless a downstream objective also requires pose agreement.

The joint view is more informative than comparing two marginal percentages alone. Equal identity and pose accuracies could coexist with many discordant references if the two kinds of disagreement offset. A small marginal gap therefore does not establish that the measures make the same claims about each query. The underlying paired events must remain available for inspection.

Agreement methodologies in Bland and Altman, 1986 and Lin, 1989 provide broader motivation for this separation. Correlation between two scores is not sufficient evidence that they can be substituted for each other. Our outcomes are binary attribute matches, so we do not apply their continuous-measure procedures directly. We use an explicit paired event table and person-level summaries instead.

The example also separates retrieval agreement from an identity-verification threshold. Every query is forced to select a known roster member; no unknown person is present. A match to the wrong roster identity is therefore a finite classification outcome, not an estimate of an operational false-accept rate for unrestricted users.

Estimating attribute differences across people

We average each attribute indicator within person, giving each of twenty identities equal weight despite minor differences in available source captures. The first regression outcome is that person's identity-agreement fraction minus pose-agreement fraction. The second is the fraction whose retrieved identity agrees while pose differs. Both are defined before fitting an intercept-only mean model.

The two models use HC3 covariance and t19 intervals, following the heteroskedasticity-robust rationale of White, 1980. Their R-squared is zero by construction; there are no controls or fixed effects. The uncertainty describes variation in the supplied person summaries conditional on this gallery and capture split, not future performance for a random sample of all possible identities.

The second quantity is unconditional on identity success: its denominator is all observed queries for a person. A conditional rate among correctly identified queries would have a different denominator and weighting. We report the joint event counts so that the reader can distinguish those interpretations instead of silently switching between them.

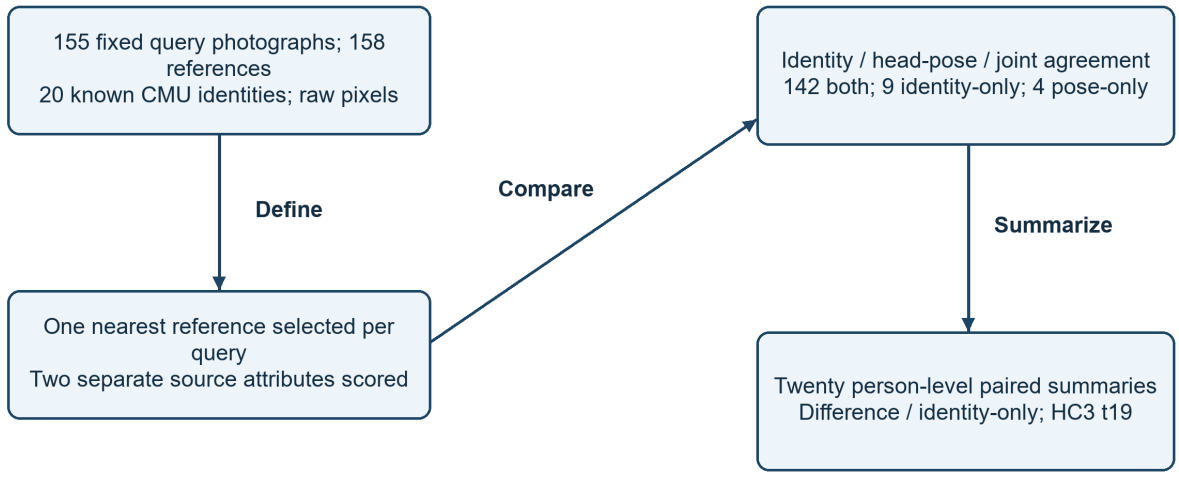

Figure 1. *One retrieval supports two explicitly different attribute assessments. The arrows describe label scoring and do not imply that pose agreement causes identity agreement or vice versa.*

Similar aggregate scores concealed thirteen discordant references

Of 155 retrieved references, 142 matched both identity and pose. Nine matched identity while carrying a different pose category. Four matched pose while depicting a different roster identity. No reference in this finite result disagreed on both attributes. That final zero is an observed count, not a guarantee that joint disagreement is impossible in other galleries.

Equal-person identity agreement was 0.975000 and pose agreement 0.943750. Their mean difference was 0.031250, with standard error 0.022551 and working interval -0.015951 to 0.078451. The marginal advantage of identity agreement was therefore uncertain under this small person-level analysis. We do not convert the interval spanning zero into a statement that the attributes are equivalent.

Identity-only agreement averaged 0.056250 across people, with standard error 0.021770 and interval 0.010684 to 0.101816. The corresponding pose-only fraction was 0.025000 descriptively. These two kinds of discordance explain why the marginal difference is smaller than the total number of references that support different claims about the query.

Table 1. *Attribute-agreement summaries for the same retrieved references. N counts twenty person means; intercept-only OLS with HC3 covariance and t19 intervals. No controls or fixed effects. Identity-only uses all queries as its denominator rather than conditioning on successful identity retrieval.*

Model	Term	Estimate (SE)	95% CI	N
M1	Mean outcome/contrast	0.031250 (0.022551)	[-0.015951, 0.078451]	20
M1: fit	R-squared = 0.0000	-	-	20
M2	Mean outcome/contrast	0.056250 (0.021770)	[0.010684, 0.101816]	20
M2: fit	R-squared = 0.0000	-	-	20

Model labels: M1 = Identity-match minus pose-match fraction; M2 = Identity-correct but pose-different fraction.

The result is therefore a construct audit rather than evidence that one score is scientifically superior. If the intended task is identity recognition across pose, the nine identity-only cases need not be penalized. If a task requires matching the view, an identity score alone cannot certify success. The objective determines the interpretation, while the measured event remains the same.

The joint table also prevents one disagreement from being hidden by another. The four pose-only matches offset part of the nine identity-only matches in the marginal difference, leaving a gap of only five query outcomes before person weighting. A reader seeing just two high percentages might miss that thirteen selected references support different attribute claims. Retaining the paired reference identifier makes that distinction directly inspectable rather than dependent on a verbal interpretation of an aggregate score.

Because both labels come from the same selected photograph, the analysis does not isolate the causal contribution of pose to identity recognition. Changing a gallery's pose coverage could change which reference is selected and therefore both scores together. The current operation holds retrieval fixed precisely to audit interpretation. It should not be confused with an intervention that controls one visual attribute while holding all others unchanged.

Related metrics answer other questions again

Tracking evaluation offers a useful parallel. Luiten et al., 2020 separates aspects of tracking quality rather than making every failure mode disappear into one undifferentiated count. Our still-image experiment does not compute tracking accuracy or temporal association, but the parallel reinforces the value of keeping distinguishable performance dimensions explicit.

Distributional generation metrics in Heusel et al., 2017 and Unterthiner et al., 2018 assess properties different from a single roster-identity decision. A collection of plausible images or videos can receive a favorable distributional score without demonstrating which subject is depicted in each item. We do not compute either metric here or infer their values from our attribute-match counts.

Huang et al., 2023 organizes video assessment into multiple dimensions, supporting the broader principle that visual outputs can satisfy one requirement while failing another. Our two labels cover only a tiny part of that space. They contain no assessment of temporal smoothness, realism, motion quality, semantic instruction following, or human acceptability.

Perceptual comparison methods such as Zhang et al., 2018, Ding et al., 2020, and Wang et al., 2004 emphasize learned similarity, structure, or texture through their respective definitions. Those quantities need not equal identity-label agreement. Their relevance is to choosing a measurement appropriate to the claim; none was executed as an alternative scorer in this controlled retrieval audit.

Semantics do not remove the need to name identity

Vision-language representations in Radford et al., 2021 and alignment evaluation in Hu et al., 2023 concern relationships between visual content and language. Ghosh et al., 2023 and Huang et al., 2023a examine compositional aspects of generated outputs. They could support useful additional assessments in a generative experiment, but they do not turn an unspecified reference similarity into a complete identity-and-pose certificate.

For example, an image can satisfy a textual request for a person facing left while depicting someone other than the intended subject. An identity-preserving image can also deliberately follow an instruction to change head pose. These are conceptual possibilities, not extra observations generated for this paper. They show why combining dimensions requires an explicit task definition rather than assuming every favorable score rewards the same behavior.

The source category itself is coarse. Two left-labeled captures can differ in actual angle and expression geometry, while two neighboring physical angles could fall on opposite category labels. We therefore avoid describing pose-label equality as precise view reconstruction. No angular tolerance or perceptual pose threshold has been validated with the acquired data.

Limits and evidence preservation

Whole-frame distance can exploit nonfacial context, and all queries belong to a small familiar roster. The image captures are not generated outputs, and the gallery contains source pose annotations that may not be available in another workflow. Source labels were accepted as the comparison definition rather than independently re-annotated. These boundaries prevent the result from establishing general identity preservation or a learned pose estimator's validity.

Documentation in Gebru et al., 2021 and reuse principles in Wilkinson et al., 2016 motivate retaining the chosen reference for every query, both attribute labels, joint events, and person-level summaries. That record makes an assessment claim reviewable. It also keeps this construct question distinct from other

uses of the same archive for missing regions, pose transfer, and reference budgets; those analyses are not independent replications of the present audit.

Conclusion

The same retrieved reference can support identity agreement without supporting pose agreement, or the reverse. Thirteen such discordant references occurred even though both aggregate agreement rates were high. A reference-based evaluation should therefore name which attribute it rewards and preserve joint outcomes when multiple claims matter. The evidence here establishes that distinction within a fixed source-label retrieval task, with no claim to have evaluated a subject-preserving generation model.

References

1. 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.
2. Radford, Alec, et al. "Learning Transferable Visual Models From Natural Language Supervision" *arXiv preprint arXiv:2103.00020*, 2021. https://arxiv.org/abs/2103.00020.
3. Meng, Qiang, et al. "MagFace: A Universal Representation for Face Recognition and Quality Assessment." *arXiv preprint arXiv:2103.06627*, 2021. https://arxiv.org/abs/2103.06627.
4. 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.
5. 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.
6. 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.
7. 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.
8. Lin, Lawrence I-Kuei. (1989). A Concordance Correlation Coefficient to Evaluate Reproducibility. Biometrics. https://doi.org/10.2307/2532051.
9. Meng, Zijie; Liu, Jiwen; Liu, Yufei; Tong, Chengzhuo; Liu, Xiaoqiang; Zhang, Yuanxing; Xu, Yulong; Wan, Pengfei. (2026). ARGUS: Stacked Multi-View Identity Mosaic Injection for Subject-Preserving Video Generation. arXiv:2606.11670. https://arxiv.org/abs/2606.11670.
10. Wilkinson, Mark D.; Dumontier, Michel; Aalbersberg, IJsbrand Jan; Appleton, Gabrielle; Axton, Myles; Baak, Arie; Blomberg, Niklas; Boiten, Jan-Willem; da Silva Santos, Luiz Bonino; Bourne, Philip E.; Bouwman, Jildau; Brookes, Anthony J.; Clark, Tim; Crosas, Mercè; Dillo, Ingrid; Dumon, Olivier; Edmunds, Scott; Evelo, Chris T.; Finkers, Richard; Gonzalez-Beltran, Alejandra; Gray, Alasdair J.G.; Groth, Paul; Goble, Carole; Grethe, Jeffrey S.; Heringa, Jaap; ’t Hoen, Peter A.C; Hooft, Rob; Kuhn, Tobias; Kok, Ruben; Kok, Joost; Lusher, Scott J.; Martone, Maryann E.; Mons, Albert; Packer, Abel L.; Persson, Bengt; Rocca-Serra, Philippe; Roos, Marco; van Schaik, Rene; Sansone, Susanna-Assunta; Schultes, Erik; Sengstag, Thierry; Slater, Ted; Strawn, George; Swertz, Morris A.; Thompson, Mark; van der Lei, Johan; van Mulligen, Erik; Velterop, Jan; Waagmeester, Andra; Wittenburg, Peter; Wolstencroft, Katherine; Zhao, Jun; Mons, Barend. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. https://doi.org/10.1038/sdata.2016.18.
11. Ding, Keyan, et al. "Image Quality Assessment: Unifying Structure and Texture Similarity." *arXiv preprint arXiv:2004.07728*, 2020. https://arxiv.org/abs/2004.07728.
12. Mitchell, Tom. (1997). CMU Face Images [Dataset]. UCI Machine Learning Repository. https://doi.org/10.24432/C5JC79.
13. Huang, Ziqi, et al. "VBench: Comprehensive Benchmark Suite for Video Generative Models" *arXiv preprint arXiv:2311.17982*, 2023. https://arxiv.org/abs/2311.17982.
14. 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.

15. Carnegie Mellon University. (2026). Neural Networks for Face Recognition: companion data and file documentation [Original educational archive; accessed8November2026]. https://www.cs.cmu.edu/~tom/faces.html.

16. 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.

17. Luiten, Jonathon, et al. "HOTA: A Higher Order Metric for Evaluating Multi-Object Tracking" *arXiv preprint arXiv:2009.07736*, 2020. https://arxiv.org/abs/2009.07736.

18. 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.

19. Carroll, Raymond J., David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. “Measurement Error in Nonlinear Models.” 2006. https://doi.org/10.1201/9781420010138.

20. Kim, Minchul, et al. "AdaFace: Quality Adaptive Margin for Face Recognition." *arXiv preprint arXiv:2204.00964*, 2022. https://arxiv.org/abs/2204.00964.

21. 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.

22. 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.

23. Stevens, S. S.. (1946). On the Theory of Scales of Measurement. Science. https://doi.org/10.1126/science.103.2684.677.

24. 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.

25. Huang, Kaiyi, et al. "T2I-CompBench++: An Enhanced and Comprehensive Benchmark for Compositional Text-to-image Generation." *arXiv preprint arXiv:2307.06350*, 2023. https://arxiv.org/abs/2307.06350.

26. 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.

27. Ghosh, Dhruba, et al. "GenEval: An Object-Focused Framework for Evaluating Text-to-Image Alignment." *arXiv preprint arXiv:2310.11513*, 2023. https://arxiv.org/abs/2310.11513.

28. Hu, Yushi, et al. "TIFA: Accurate and Interpretable Text-to-Image Faithfulness Evaluation with Question Answering." *arXiv preprint arXiv:2303.11897*, 2023. https://arxiv.org/abs/2303.11897.

29. Gebru, Timnit, et al. "Datasheets for datasets" *Communications of the ACM*, 2021. https://doi.org/10.1145/3458723.

30. Heusel, Martin, et al. "GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium" *arXiv preprint arXiv:1706.08500*, 2017. https://arxiv.org/abs/1706.08500.