Representation Audits without Inferring Demographics from Faces

Lucas Allen (AT), Scott Butler (IS), Elijah Reed (CI)

Abstract

An identity benchmark can be audited using documented sampling and observation records without assigning demographic categories from facial appearance. This article specifies what such an audit can establish and where its evidence ends. ARGUS motivates subject-preserving video evaluation. No demographic labels are inferred and no benchmark images are processed. The proposed audit distinguishes identity frequency, reference coverage, and appropriately documented population attributes, preventing an accessible metadata analysis from becoming an unsupported claim about demographic representation or fairness.

Counting observations is a narrower task than classifying people

A reference collection contains identities, images, views, and repeated observations. Their frequencies can reveal concentration and unequal evidence budgets. They cannot by themselves establish the demographic composition of the people represented. Inferring categories from names, appearance, or public familiarity adds a new measurement process with assumptions that the original metadata may not support. An audit should name the observations it actually has rather than silently replace missing population information with an impression.

ARGUS uses structured identity evidence for video generation (Meng et al., 2026b). This makes reference representation a relevant evaluation concern, but it does not provide every annotation needed for a population fairness claim. The present article proposes a metadata-first audit and performs no demographic classification. Its objective is to separate readily verifiable properties of a benchmark's organization from stronger claims about people, groups, or generated outcomes that would require appropriate additional evidence.

What the available records can support

A proposed audit would inventory documented subject identifiers, image counts, reference reuse, view labels, source sessions, and missing records. It could describe how much of the collection is concentrated in frequently observed subjects and how reference budgets vary. These are properties of the recorded benchmark. The audit should preserve uncertainty in subject linkage and distinguish an image copy from a genuinely new observation where provenance allows that distinction.

If demographic annotations already exist, the audit would first inspect their definitions, source, collection purpose, and uncertainty. An available label is not automatically suitable for every research question. Categories can refer to different constructs, and a label assigned for one purpose may not support a conclusion about another. The analysis should use only attributes whose provenance and interpretation fit the stated task, while describing absent information as a limit rather than imputing it from visual resemblance.

Coverage and outcome evidence should remain separate. A balanced count under an appropriate annotation scheme does not establish equal generation performance. An uneven count does not identify the direction or magnitude of any performance disparity. A follow-up outcome study would require assessable generated clips, valid measurements, sufficient observations, and an explicit comparison. The metadata audit can guide that study's sampling but cannot supply its unobserved identity errors through a statistical adjustment.

Connecting representation to a defensible evaluation

The unit of representation should match the claim. Equal numbers of images can hide unequal numbers of subjects, and equal subject counts can hide different view coverage. A useful report would show these dimensions separately under the available documentation. It should also state the benchmark's source population and selection process. A convenience collection remains a convenience collection after reweighting; weights can change emphasis among observed records but cannot create people or conditions that were never included.

For generated outcomes, the evaluator's own coverage becomes relevant. A score unavailable for particular observed conditions can create selective missingness, although explaining its population pattern requires suitable labels and evidence. The proposed audit would retain unassessable outputs as a separate outcome. It would not infer demographic mechanisms from their appearance. That boundary avoids turning a potentially useful observation about measurement coverage into a claim about group performance that the data cannot substantiate.

Documentation should describe which questions were deliberately left unanswered and which additional records would be required. This is a constructive research output rather than an empty disclaimer. It allows a later study to obtain appropriate annotations and design its comparisons before inspecting outcome differences. Versioned records also make changes in benchmark composition visible, preventing a revised collection from being compared with an earlier one as though the same subjects, views, and assessment opportunities had remained fixed throughout.

What can be counted without visual inference

For the proposed comparison, the key observation is documented subject identifier. A reference collection is rich in views of a few subjects but lacks appropriate population attribute annotations. This hypothetical case poses a question about concentration in known observation units. Record distribution of source images and distribution of documented subject and view coverage separately. A visually diverse contact sheet cannot establish an identified demographic distribution. Record what the metadata can describe and separate any proposed outcome audit from unobserved attributes.

Count the number of reference views per documented identity and record how much of the collection comes from repeated acquisitions. This supports an audit of concentration and observable coverage. If demographic categories are unavailable, list the precise attributes missing from the inventory. A future fairness comparison would need appropriately documented attributes and outcomes for a justified task. Changes in weighting among known identities can improve representation of the observed sources. They cannot recover information about people and groups absent from the record.

Set the comparison rules for distribution of source images in advance and explain how the assessment would represent failures to generate an observation. For the documented subject identifier, keep undetectable faces in the log and state which cases remained eligible for recognition.

Store conditioning views, generated frames and visibility annotations separately for documented subject identifier. To evaluate concentration in known observation units, keep paired observations together during resampling, since independent resampling would lose their shared source of variation. When examining distribution of documented subject and view coverage, inspect pairs of frames with similar pose before drawing conclusions about a persistent identity error.

For concentration in known observation units, define one setting for the primary estimate and show the neighboring settings as a separate comparison. The evaluation of distribution of source images can

distinguish an incorrect subject from a permitted change in expression, pose or scene. Collect a separate set of documented subject identifier under documented conditions before claiming that the contrast generalizes.

Samples and evidence

An existing collection provides a starting point for auditing subgroup coverage while leaving the deployment distribution to be specified separately. One study defines an unconstrained face verification collection and pair protocol (Huang et al., 2008). One approach constructs a face attribute dataset with demographic balance (Kärkkäinen and Joo, 2019). In practical terms, pair definitions determine the verification denominator, so any comparison should state how multiple views of the same person were sampled.

Analysis and interpretation

Repetitions from a common generating condition can have correlated errors, making the number of independent groups more informative than the nominal row count. Single-reference conditioning provides a budget-matched comparison for identifying which errors disappear when extra views are made available to the generator. Earlier research conditions identity-preserving image synthesis on a single reference (Wang et al., 2024). In a controlled test, one approach provides guidance on inference under dependence within specified groups (Cameron and Miller, 2015).

Uncertainty intervals and performance profiles help show whether an apparent advantage persists across titles with very different return distributions. One published study describes estimation and implementation of linear mixed-effects models (Bates et al., 2015). For a prospective comparison, mixed-effects models account for repeated observations of the same unit, while the choice of grouping still needs a design rationale. One approach investigates reliable comparisons of reinforcement learning under limited runs (Agarwal et al., 2021).

Matching can improve comparability on recorded attributes; an overlap display should identify which observations are lost before the outcome contrast is computed. A family of endpoints should be declared before evaluation so that reported discoveries reflect every comparison eligible to appear in the result. In a controlled test, a relevant contribution controls the false discovery rate for a declared family of hypotheses (Benjamini and Hochberg, 1995). One approach reviews matching as a design step for comparing observational units (Stuart, 2010).

A stacked portrait input permits a test of view coverage, provided the evaluator retains an independent portrait and records near-duplicate input views. When diagnostic endpoints are dependent, the correction should reflect that dependence and use a defined family rather than whichever comparisons looked notable. For implementation, a methodological precedent builds a stacked identity representation from input portraits (Li et al., 2023a). One approach examines false discovery rate control when tests are dependent (Benjamini and Yekutieli, 2001).

Systems and comparators

A one-reference adapter supplies a baseline for evaluating whether added views help temporal persistence when motion and instructions are held constant. For implementation, record inputs, preprocessing and model versions to locate failures caused by dependencies outside the model. A methodological precedent personalizes video through a face adapter using one reference (He et al., 2024). Related work examines maintenance costs from machine learning dependencies (Sculley et al., 2015).

One published study adds image prompting to a text-conditioned diffusion model (Ye et al., 2023). In practical terms, a declared software pipeline can preserve the order of transformations and enable held-out validation without refitting preprocessing on the evaluation observations. Varying the image-conditioning strength gives a practical control for the trade between copying a reference and following the requested transformation. One approach describes reusable software interfaces for fitting and evaluating estimators (Pedregosa et al., 2011).

A learned concept embedding offers a compact control for comparing changes in prompt adherence under the same set of reference images. One published study adapts image synthesis to a specific subject using example images (Ruiz et al., 2022). In an evaluation, a relevant contribution learns concept embeddings from a few images (Gal et al., 2022). Keep reference and evaluation views separate so that recognition after adaptation can be distinguished from successful reproduction of a training portrait.

Measurement and error

Related work uses an additive angular margin to train discriminative face embeddings (Deng et al., 2018). A methodological precedent investigates statistical consequences of imperfectly measured variables (Carroll et al., 2006). In a controlled test, separating uncertainty in the measured inputs from variability in the underlying property is important when the same noisy proxy also defines the strata. Angular discrimination makes a reasonable verification baseline, provided the threshold is evaluated again on profiles, occlusions and synthetic views outside ordinary portrait capture.

A relevant contribution provides object annotations and standardized visual recognition tasks (Lin et al., 2014). For implementation, landmark localization provides a way to identify crop failures before measuring resemblance and to retain a denominator for faces that were never detected. Object annotations support localized checks, with the class mapping and annotation policy recorded before evaluating cases absent from the detector vocabulary. One approach jointly localizes faces and facial landmarks (Deng et al., 2019).

Another study uses a cosine margin for discriminative face recognition (Wang et al., 2018). One published study reports prediction performance jointly with abstention coverage (Geifman and El-Yaniv, 2017). During validation, a cosine-margin baseline helps distinguish rejection due to the representation from rejection caused by the threshold. Report coverage beside error rates. Rejecting difficult cases can improve conditional accuracy while leaving fewer cases with an answer.

Linking evidence to the evaluation

The size and makeup of the distractor gallery alter the false-match burden, so gallery construction belongs in the main experimental description. Serial dependence makes neighboring measurements count differently for uncertainty, even if the average contrast itself is calculated using all observed time points. In practical terms, one approach tests face recognition with a large distractor collection (Kemelmacher-Shlizerman et al., 2015). A methodological precedent allows covariance estimates to account for heteroskedasticity and serial dependence (Newey and West, 1987).

One approach specifies principles for discoverable and reusable research artifacts (Wilkinson et al., 2016). One published study asks dataset creators to record collection, intended use and limitations (Gebru et al., 2021). One study provides face observations with substantial variation in pose and age (Cao et al., 2017). In a controlled test, pose and age diversity help construct cross-condition verification pairs, but transfer tests still need outputs from the intended workflow. A reproducible result requires identifiers for both the released inputs and the exact intermediate tables from which the paper derives its figures. Documenting

construction and intended use clarifies which exclusions occurred during collection, which arose from a model failure and which were imposed during analysis.

A multiple-view conditioning baseline is most informative when the views add distinct pose evidence rather than repeated copies of a similar crop. Adaptation comparisons can report trainable parameters together with sample count and compute, keeping the data budget distinct from the architectural change. One study adapts pretrained models through low-rank parameter updates (Hu et al., 2022). For a prospective comparison, an existing method uses auxiliary views to condition identity-preserving video (Wu et al., 2026).

Conclusion

A representation audit can establish concentration, observation coverage, and documented selection without inferring demographic categories from faces. This article proposes that bounded approach and reports no new population or fairness result. Its contribution is to connect every claim to an appropriate record: metadata describes benchmark organization, suitable annotations describe specified population attributes, and independently assessed outputs are required before differences in subject-preservation performance can be established.

References

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

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.

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.

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.

Meng, Zijie, et al. "Argus: Stacked multi-view identity mosaic injection for subject-preserving video generation." *arXiv preprint arXiv:2606.11670* (2026).

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.

Stuart, Elizabeth A. "Matching Methods for Causal Inference: A Review and a Look Forward." *Statistical Science*, vol. 25, no. 1, 2010, pp. 1-21. https://doi.org/10.1214/09-sts313.

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.

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.

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.

Gal, Rinon, et al. "An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion." *arXiv preprint arXiv:2208.01618*, 2022. https://arxiv.org/abs/2208.01618.

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.

Hu, Edward J., et al. “LoRA: Low-Rank Adaptation of Large Language Models.” *International Conference on Learning Representations*, 2022. https://openreview.net/forum?id=nZeVKeeFYf9.

He, Xuanhua, et al. "ID-Animator: Zero-Shot Identity-Preserving Human Video Generation." *arXiv preprint arXiv:2404.15275*, 2024. https://arxiv.org/abs/2404.15275.

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

Wang, Qixun, et al. "InstantID: Zero-shot Identity-Preserving Generation in Seconds." *arXiv preprint arXiv:2401.07519*, 2024. https://arxiv.org/abs/2401.07519.

Agarwal, Rishabh, et al.. "Deep Reinforcement Learning at the Edge of the Statistical Precipice" *arXiv preprint arXiv:2108.13264*, 2021. https://arxiv.org/abs/2108.13264.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker. “Fitting Linear Mixed-Effects Models Using lme4.” *Journal of Statistical Software*, vol. 67, no. 1, 2015. https://doi.org/10.18637/jss.v067.i01.

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

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.

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.

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.

Deng, Jiankang, et al. "RetinaFace: Single-stage Dense Face Localisation in the Wild." *arXiv preprint arXiv:1905.00641*, 2019. https://arxiv.org/abs/1905.00641.

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.

Benjamini, Yoav and Daniel Yekutieli. “The control of the false discovery rate in multiple testing under dependency.” *The Annals of Statistics*, vol. 29, no. 4, 2001, pp. 1165-1188. https://doi.org/10.1214/aos/1013699998.

Li, Zhen, et al. "PhotoMaker: Customizing Realistic Human Photos via Stacked ID Embedding." *arXiv preprint arXiv:2312.04461*, 2023. https://arxiv.org/abs/2312.04461.

Newey, Whitney K., and Kenneth D. West. “A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix.” *Econometrica*, vol. 55, no. 3, 1987, pp. 703-708. https://doi.org/10.2307/1913610.

Lin, Tsung-Yi, et al. “Microsoft COCO: Common Objects in Context.” *Lecture Notes in Computer Science*, 2014, pp. 740-755. https://doi.org/10.1007/978-3-319-10602-1_48.

Wu, Mingyang, et al. "ConsID-Gen: View-Consistent and Identity-Preserving Image-to-Video Generation." *arXiv preprint arXiv:2602.10113*, 2026. https://arxiv.org/abs/2602.10113.

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.