Occlusion, Accessories, and Missing Evidence in Identity Preservation

Leo Harper (SZ), Trevor Simmons (UY), Gabriel Elliott (VE)

Abstract

An occluded face and a changed accessory can remove different kinds of evidence from a subject-preservation test. This conceptual study proposes separating physical visibility, reference availability, and the evaluator's ability to identify the person. ARGUS motivates the question through its emphasis on robust identity conditioning. No occlusion experiment is performed here. The proposed record distinguishes a generator that loses the subject from an instrument that cannot assess the visible output, while preventing accessories or surrounding context from silently becoming the sole definition of identity.

Missing pixels are not one uniform condition

A face can be partly hidden by an object, cropped by the image boundary, turned away, or covered by an accessory. These conditions differ in what is visible and in what the task expects to remain inferable. They can also be confused in a generic difficulty label. A benchmark should state which information is removed, whether the occluder is part of the requested scene, and what other evidence remains available to the generator and evaluator.

ARGUS uses structured identity references and motivates robustness under difficult observations (Meng et al., 2026b). That does not establish that every form of missing evidence is equivalent or that any failure has one common cause. This article proposes a conceptual decomposition rather than a new test of the model. It performs no image manipulation, recognition analysis, or generated-video measurement. The proposed distinctions are intended to make a later occlusion study specific enough to support or contradict a preservation claim.

Three records that should not be collapsed

The first record concerns visible output evidence: which regions or identity-relevant observations are available at a particular time. The second concerns the reference set: whether an unoccluded or alternative view was supplied. The third concerns measurement: whether the chosen instrument can assess identity under the resulting condition. A clip can be hard for one instrument while containing enough evidence for a human observer who sees the preceding history. Conversely, a detector can produce a confident score from a small region without establishing the intended person to an independent observer.

Accessories create an additional task distinction. Keeping a requested hat or glasses can be correct, while copying an unwanted accessory from a reference can violate an instruction. The accessory can also become a powerful matching cue. A useful protocol would therefore specify whether it is a required persistent object, a permitted variation, or an intended change. The same pixel-level similarity can have different meanings under those contracts. The evaluator needs the contract before it interprets the output.

The study could vary a nominated occlusion while preserving other conditions as closely as practical. However, creating an occlusion can change lighting, motion, or image boundaries. Those consequences should be documented rather than assumed absent. An artificial mask is not automatically equivalent to a naturally moving occluder. The design should say which population it represents and avoid generalizing from one convenient manipulation to every kind of missing identity evidence.

Preserving a meaningful denominator over time

A video includes intervals with different visibility. Averaging identity scores only over detectable frames can conceal a brief but consequential subject substitution or systematically omit the hardest period. A report should retain the requested time window, detection availability, valid judgments, and any deferrals. It should distinguish an uncertain identity verdict from evidence that the subject changed. Those are different outcomes and may require different follow-up observations.

Temporal context can legitimately improve assessment. A person may remain recognizable through a brief occlusion because the surrounding trajectory is coherent. A frame-only instrument and a sequence-level observer therefore answer different questions. A benchmark should report which information each receives and avoid comparing their numbers as though both operated on the same evidence. A direct test could assess the same outputs under several declared context windows, without attributing every difference to the generator.

Reference support also needs matched comparison. A method that receives an informative unobstructed view may solve a different problem from one conditioned only on a difficult first frame. Both can be useful workflows. The evaluation should report their information and resource budgets instead of treating every resulting score difference as an isolated architectural effect. A robust preservation claim requires the evidence available to the system and the evidence available to its judge to remain recoverable.

Different masks create different missing observations

For the proposed comparison, the key observation is occlusion event. A hand briefly blocks the mouth while sunglasses obscure the eyes throughout a second clip. This hypothetical case poses a question about what evidence supports identity before and after each mask. Record temporary hand occlusion and persistent accessory occlusion separately. A detector can fail on glasses glare even when an independent viewer can follow the person. Preserve exact event times and compare automatic availability with reviewer visibility at the same intervals.

Define separate time windows for the transient hand cover and the persistent accessory. In the first case, the critical observation may be the recognizable face immediately after removal. In the second, face evidence might remain incomplete throughout the clip. A detector's failed crop and a reviewer's uncertain identity judgment are different events and can overlap. Record both along with the available nonface context. Treat a changed person after the occluder disappears as a different outcome from an interval that could never be evaluated.

Set the comparison rules for temporary hand occlusion in advance and explain how the assessment would represent failures to generate an observation. For the occlusion event, keep undetectable faces in the log and state which cases remained eligible for recognition.

Store conditioning views, generated frames and visibility annotations separately for occlusion event. If there are many observations of occlusion event, decide which ones share an underlying history and keep that grouping in any uncertainty analysis. When examining persistent accessory occlusion, inspect pairs of frames with similar pose before drawing conclusions about a persistent identity error.

After evaluating persistent accessory occlusion, list the cases whose result changes when the measurement definition is adjusted within a declared range. The evaluation of temporary hand occlusion can distinguish an incorrect subject from a permitted change in expression, pose or scene. Collect a separate set of occlusion event under documented conditions before claiming that the contrast generalizes.

Measurement and error

An existing method uses feature norms to adapt training margins to face image quality (Kim et al., 2022). A methodological precedent jointly localizes faces and facial landmarks (Deng et al., 2019). During validation, landmark localization provides a way to identify crop failures before measuring resemblance and to retain a denominator for faces that were never detected. Feature magnitude and recognition confidence can help identify low-quality crops, but agreement with an independent quality annotation should be checked first.

A relevant contribution learns facial representations with a quality-related feature magnitude (Meng et al., 2021). Report coverage beside error rates. Rejecting difficult cases can improve conditional accuracy while leaving fewer cases with an answer. In practical terms, earlier research reports prediction performance jointly with abstention coverage (Geifman and El-Yaniv, 2017). A quality-sensitive representation can be used to define inspectable strata before comparing acceptance rates across views with very different visibility.

A relevant contribution provides object annotations and standardized visual recognition tasks (Lin et al., 2014). During validation, a multiscale score can help distinguish sensitivity to resampling or crop size from a stable image-level judgment. Related work uses multiscale image representations for quality prediction (Ke et al., 2021). Object annotations support localized checks, with the class mapping and annotation policy recorded before evaluating cases absent from the detector vocabulary.

A whole-image embedding reflects scene description and background alongside subject appearance. A local check can distinguish these contributions. Report the detection and association components as well as the aggregate so that tracking continuity can be interpreted alongside object availability. In a controlled test, one study learns aligned visual and language representations through contrastive training (Radford et al., 2021). One study combines detection and association considerations in tracking assessment (Luiten et al., 2020).

Verification should include genuine and impostor pairs with similar acquisition conditions to avoid turning image quality into a surrogate for identity. A relevant contribution contrasts precision-recall and receiver operating characteristic summaries under class imbalance (Saito and Rehmsmeier, 2015). Low prevalence can make a seemingly small false-positive rate operationally costly, so report precision at plausible recall levels and retain class counts. One approach learns an embedding space for face verification and recognition (Schroff et al., 2015).

Confidence calibrated on the original collection may change under a new capture or generation regime, so calibration belongs in transfer testing. In a controlled test, related work compares choices of forecast error scale across series (Hyndman and Koehler, 2006). Scaling choices determine how units of different magnitude contribute to an aggregate error; a transparent report retains the original-scale results where interpretable. Earlier research examines uncertainty estimates under dataset shift (Ovadia et al., 2019).

Another study investigates statistical consequences of imperfectly measured variables (Carroll et al., 2006). Earlier research formulates classification decisions with explicit costs for different errors (Elkan, 2001). For implementation, separating uncertainty in the measured inputs from variability in the underlying property is important when the same noisy proxy also defines the strata. An application with asymmetric errors calls for costs to be stated separately from the metric so that operating thresholds can be recalculated.

One published study uses an additive angular margin to train discriminative face embeddings (Deng et al., 2018). Another study assesses probability calibration and studies temperature scaling (Guo et al., 2017).

For implementation, after fitting a decision threshold, reliability curves on an independent split can show whether reported confidence still matches observed success frequency. Angular discrimination makes a reasonable verification baseline, provided the threshold is evaluated again on profiles, occlusions and synthetic views outside ordinary portrait capture.

Report which raters, measurements and agreement definition enter the coefficient, since different designs give superficially similar statistics different meanings. Separate structure and texture responses are informative when a change in surface detail is permissible but a change in overall shape is consequential. A relevant contribution distinguishes forms of intraclass correlation according to the rating design (Koo and Li, 2016). Earlier research constructs a full-reference metric with structure and texture components (Ding et al., 2020).

Systems and comparators

One study describes reusable software interfaces for fitting and evaluating estimators (Pedregosa et al., 2011). Use a text-conditioned video baseline to assess generic motion with the same clip duration and language instructions. An existing method uses a diffusion transformer for text-conditioned video generation (Yang et al., 2024). For a prospective comparison, a declared software pipeline can preserve the order of transformations and enable held-out validation without refitting preprocessing on the evaluation observations.

Analysis and interpretation

A missingness analysis distinguishes failures from observations that were never eligible and makes the assumptions behind complete-case summaries visible to readers. Serial dependence makes neighboring measurements count differently for uncertainty, even if the average contrast itself is calculated using all observed time points. Related work allows covariance estimates to account for heteroskedasticity and serial dependence (Newey and West, 1987). In a controlled test, one approach separates missing-data assumptions from observed-data summaries (Little and Rubin, 2002).

One published study describes estimation and implementation of linear mixed-effects models (Bates et al., 2015). In an evaluation, matching can improve comparability on recorded attributes; an overlap display should identify which observations are lost before the outcome contrast is computed. Mixed-effects models account for repeated observations of the same unit, while the choice of grouping still needs a design rationale. A methodological precedent reviews matching as a design step for comparing observational units (Stuart, 2010).

Samples and evidence

Pose and age diversity help construct cross-condition verification pairs, but transfer tests still need outputs from the intended workflow. In practical terms, another study constructs a face attribute dataset with demographic balance (Kärkkäinen and Joo, 2019). One study provides face observations with substantial variation in pose and age (Cao et al., 2017). An existing collection provides a starting point for auditing subgroup coverage while leaving the deployment distribution to be specified separately.

Linking evidence to the evaluation

One approach examines discrepancies between paired measurement procedures (Bland and Altman, 1986). A methodological precedent uses appearance information in online association (Wojke et al., 2017). Related work defines an unconstrained face verification collection and pair protocol (Huang et al., 2008). A paired difference plot can reveal systematic offsets and score-dependent disagreement even when the correlation between two procedures is high. An appearance-aware baseline can be compared with motion-only association under the same detections to locate the source of identity changes. During validation, pair definitions determine the verification denominator, so any comparison should state how multiple views of the same person were sampled.

Conclusion

Occlusion assessment should distinguish missing output evidence, missing reference evidence, and missing measurement capability. Accessories and context also need an explicit task role. The proposed design retains time and coverage rather than silently scoring only easy visible frames. It reports no new experiment or failure rate. Its contribution is a framework for obtaining a bounded preservation result in which a generator's behavior is not confused with the limits of the evidence or the evaluator used to judge it.

References

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.

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

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.

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.

Ke, Junjie, et al.. "MUSIQ: Multi-scale Image Quality Transformer" *arXiv preprint arXiv:2108.05997*, 2021. https://arxiv.org/abs/2108.05997.

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.

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.

Radford, Alec, et al.. "Learning Transferable Visual Models From Natural Language Supervision" *arXiv preprint arXiv:2103.00020*, 2021. https://arxiv.org/abs/2103.00020.

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.

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.

Yang, Zhuoyi, et al. "CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer." *arXiv preprint arXiv:2408.06072*, 2024. https://arxiv.org/abs/2408.06072.

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.

Saito, Takaya and Marc Rehmsmeier. “The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets.” *PLOS ONE*, vol. 10, no. 3, 2015, pp. e0118432. https://doi.org/10.1371/journal.pone.0118432.

Hyndman, Rob J. and Anne B. Koehler. “Another look at measures of forecast accuracy.” *International Journal of Forecasting*, vol. 22, no. 4, 2006, pp. 679-688. https://doi.org/10.1016/j.ijforecast.2006.03.001.

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.

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.

Elkan, Charles. “The Foundations of Cost-Sensitive Learning.” *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence*, 2001, pp. 973-978. https://cseweb.ucsd.edu/~elkan/rescale.pdf.

Wojke, Nicolai, et al. "Simple Online and Realtime Tracking with a Deep Association Metric." *arXiv preprint arXiv:1703.07402*, 2017. https://arxiv.org/abs/1703.07402.

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

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

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.

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.

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.

Little, Roderick J. A., and Donald B. Rubin. *Statistical Analysis with Missing Data*. Wiley, 2002. https://doi.org/10.1002/9781119013563.

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.

Koo, Terry K., and Mae Y. Li. "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research." *Journal of Chiropractic Medicine*, vol. 15, no. 2, 2016, pp. 155-163. https://doi.org/10.1016/j.jcm.2016.02.012.

Ding, Keyan, et al. "Image Quality Assessment: Unifying Structure and Texture Similarity." *arXiv preprint arXiv:2004.07728*, 2020. https://arxiv.org/abs/2004.07728.

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.

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.

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.