Robustness to Corrupted and Irrelevant Identity References

Joshua Marshall (IE), Adrian Franklin (VU), Jeremy Sullivan (AD)

Abstract

Reference corruption and reference irrelevance create different challenges for a subject-preserving generator. One degrades evidence about the intended subject; the other introduces evidence that may concern another subject or an unrelated scene. This article proposes a controlled comparison that keeps those interventions distinct. ARGUS motivates organized multi-view conditioning. No references are altered and no robustness results are reported. The design records intervention severity, selection behavior, and assessment coverage so that a future score change can be connected to the actual reference problem introduced.

1. Two ways for reference evidence to become unreliable

A blurred portrait may still depict the intended person, while a sharp but unrelated portrait may depict someone else. Their visual quality and semantic relevance differ. Pooling them under a general noise label can hide whether a system handles missing detail, contradictory evidence, or reference selection. Robustness should therefore be defined in relation to a specified perturbation and task rather than inferred from performance on an unspecified mixture of difficult inputs.

ARGUS uses organized identity observations and a reference-management setting (Meng et al., 2026b). This motivates asking how a workflow handles unreliable evidence without assuming that its documented mechanism solves every corruption or conflict. The present article outlines a robustness experiment and reports no modified-reference outcomes. Its question is how to separate degradation of useful information from introduction of irrelevant information, and how to preserve that distinction through generation, scoring, and final interpretation.

Identity evidence has a sampling structure before any recognition score is calculated. LFW defines a particular pair-verification protocol (Huang et al., 2008), and the documented scikit-learn loader makes its prescribed files accessible through a reproducible software interface (Pedregosa et al., 2011). Neither source turns a pair count into a count of independent subjects or images. Dataset documentation should identify how references, identities, comparisons, and splits were assembled (Gebru et al., 2021). For generated video, the analogous record should distinguish a new person from another reference photograph, another trajectory, or another frame. Each adds evidence along a different dimension. A transparent report should preserve those links rather than flattening every observation into one score table. The resulting metadata can support a useful audit even when no images or generated clips are available. Such an audit describes coverage and organization; it does not itself establish recognition accuracy, demographic representativeness, or identity preservation in an unobserved video.

A large collection of measurements does not guarantee a large collection of independent cases. Repeated observations can share unrecorded acquisition conditions or a common source, making row-level uncertainty misleading. White's covariance estimator addresses heteroskedasticity within a specified regression (White, 1980), while cluster methods address a declared form of grouped dependence (Cameron and Miller, 2015). Neither approach repairs a poorly defined outcome or establishes representativeness of the observed sample. A report should identify the number of groups, their relative sizes, and the reason for selecting the grouping level. When only a few independent groups are available, apparently precise numerical intervals warrant particular caution. This is a limitation of the evidence structure, rather than a reason to count correlated measurements as fresh observations.

2. Designing interventions with recoverable meanings

A proposed corruption study would preserve source images and apply declared transformations at recorded severities. Each transformation can change several visual properties, so the report should avoid assigning a unique mechanism without additional observations. An irrelevance study would instead specify how unrelated references are selected and inserted. It should distinguish wrong-subject images from same-subject images that are merely unhelpful for the requested view. Those conditions provide different evidence to a selection or conditioning system.

The number of useful references and the total reference budget should be tracked independently. Replacing a useful image with an irrelevant image changes both the amount of relevant evidence and the opportunity for interference. Adding an irrelevant image changes the total budget. A design can include both comparisons, but their interpretations should remain separate. Matching only the final tile count cannot isolate the effect of contradiction from the loss of a useful observation.

The workflow's selection behavior should be recorded where its interface permits it. A system may avoid an irrelevant input rather than remain unaffected after conditioning on it. That is a legitimate workflow response, but it differs from robustness of the generation stage to the same received evidence. An end-to-end result can be reported alongside a controlled-input comparison. The analysis should not claim an isolated component mechanism when only the full pipeline has been observed.

An identity score is produced by a measurement system with its own training and assumptions. FaceNet (Schroff et al., 2015), ArcFace (Deng et al., 2018), and CosFace (Wang et al., 2018) offer influential approaches to discriminative face representation. Their outputs should be interpreted in relation to the images and decisions on which they are assessed. VGGFace2 broadens variation within subject observations (Cao et al., 2017), MegaFace makes large distractor populations important (Kemelmacher-Shlizerman et al., 2015), and FairFace addresses imbalance across annotated demographic categories (Kärkkäinen and Joo, 2019). These are complementary evaluation concerns. A generated-video study should not compress them into an unexplained claim that a face benchmark is difficult. It should specify whether difficulty comes from viewpoint, acquisition, distractors, subject population, or the representation itself. Demographic conclusions require appropriate annotations and cannot be inferred merely from identity frequencies or from a generic face-similarity score. A metric's usefulness for one purpose remains an empirical question for another.

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 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.

3. Assessing degradation, recovery, and unresolved cases

Outcomes should include recognizable identity, instruction execution, generation failures, and measurement coverage. A strong generic reference similarity can be misleading when the perturbed collection contains a wrong subject. The evaluator therefore needs a declared target reference distinct from the intervention set where appropriate. Independent judgments can establish whether a generated person follows the intended subject or the irrelevant reference, with uncertainty retained for genuinely ambiguous outputs.

Repeated perturbations and generation samples should remain linked to the original subject and source observations. Many altered versions of one image do not create many independent subjects. The study should identify whether uncertainty concerns sampled people, intervention instances, or stochastic generation. It should also retain all attempts under a matched budget. Selecting the best recovered output after many retries evaluates a recovery workflow whose additional effort must be visible in the comparison.

The final claim would identify the corruption families, relevance conditions, severity range, and reference coverage actually tested. Success under blur does not establish success under wrong-subject insertion, and success with many useful references may not transfer to a nearly empty evidence set. A useful robustness profile can remain specific. Its boundaries guide the next experiment and prevent a collection of favorable perturbation examples from becoming an unqualified assertion of resilience to arbitrary reference problems.

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.

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. Designing an interpretable preservation study

Subject preservation is one dimension of a larger video-generation task. VBench explicitly motivates separating evaluation dimensions (Huang et al., 2023), and the visual literature offers structural (Wang et al., 2004), perceptual (Zhang et al., 2018), and temporal-distribution measurements (Unterthiner et al., 2018). A study should choose among them according to the claim and available reference information. When identity, motion, and visual quality move in different directions, the disagreement should be analyzed rather than automatically resolved by a weighted total. The weighting itself expresses a preference that may differ across applications. Preserving per-frame and per-clip records allows alternative interpretations to be evaluated later. It also makes coverage visible when some poses or occlusions prevent valid scoring. A generator can demonstrate a useful, limited preservation capability without having established universal reliability across all subjects, durations, and visual conditions.

Difficult cases can disappear at several stages: data eligibility, generation, measurement, and final scoring. Each exclusion changes the population represented by the reported result. Coverage-aware evaluation makes those decisions visible (Geifman and El-Yaniv, 2017). A confidence score should also be checked when the input distribution changes, because uncertainty under shift is an empirical property rather than a guarantee of the model interface (Ovadia et al., 2019). An informative record preserves failures to produce an output and failures to obtain a valid measurement. Treating either as absent data without explanation can inflate apparent reliability. A proposed method should state its unsupported conditions and test the boundary between accepted and rejected cases as carefully as the quality of its successful outputs.

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

Reference robustness needs interventions that distinguish degraded evidence, lost evidence, and irrelevant evidence. The proposed protocol records source inputs, reference budgets, selection responses, and independent output assessments. It reports no new robustness result. Its contribution is an interpretable comparison in which a future performance change can be connected to the documented reference intervention and workflow stage, while conditions outside the tested perturbations remain open empirical questions.

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.