Expression Change without Identity Substitution in Generated Video

David Armstrong (ZA), Eric Walters (EE), Simon Webb (LY)

Abstract

An expression change can alter facial appearance without changing the intended person. This conceptual study proposes a subject-preservation evaluation that records the requested expression, the available reference expressions, and independent identity judgments. It separates a failure to animate from a failure to preserve the subject and considers measurement ambiguity during motion. ARGUS motivates the multi-view conditioning problem. No new expression labels or video outputs are analyzed. The contribution is a design for testing flexibility and preservation together, rather than rewarding a generator for keeping a familiar neutral face unchanged.

1. The target is an invariant through a permitted change

Identity preservation is often easiest to demonstrate when the generated face remains close to a familiar reference. That observation says less about a task in which the subject must speak, smile, or change expression. Those changes affect the visible face and can alter an automated similarity score. The evaluation needs to distinguish the intended deformation from a substitution of the person. Otherwise, a generator may receive a favorable preservation score for declining to perform the requested motion.

ARGUS provides a framework for conditioning video on structured identity evidence (Meng et al., 2026b). Its motivating concern includes maintaining a subject under changes in observation. This article does not claim that any particular expression failure has been observed in the system. It proposes a paired evaluation of expression execution and identity preservation. No generated clips, face embeddings, or new annotations enter the present analysis, so all experimental outcomes remain to be measured.

A benchmark's observation count should be accompanied by a description of what an observation is. The LFW pair protocol separates identities from their images and from the comparisons used for verification (Huang et al., 2008). Its documented software access routes also distinguish the ten-fold assessment from development subsets (Pedregosa et al., 2011). Dataset documentation connects those files to their collection and intended purpose (Gebru et al., 2021). A generation benchmark needs an equally explicit account of how a referenced subject becomes a set of output clips and scored frames. Without it, repeated observations may receive unintended weight or cross a supposed assessment boundary. A metadata analysis can establish whether those links are consistent. It should still avoid claiming that consistency of filenames proves recognizability, fair representation, or robustness under viewpoint change. Those are further outcomes requiring appropriate measurements.

2. The evidence needed to interpret an expression test

A reference record should identify which expressions and views are already present. A system supplied with the target expression faces a different conditioning problem from one required to produce it without that observation. Both workflows may be useful, but their information access should be stated. A comparison should match reference opportunities where possible or present the difference explicitly. It should not attribute an advantage from additional subject evidence solely to the generation mechanism.

The expression endpoint should also be defined separately from identity. It could concern whether a requested category or temporal pattern is recognizable to appropriately instructed observers. The identity endpoint would concern whether the output remains the intended subject. A clip might satisfy either one while failing the other. The study should retain the full outcome profile and avoid selecting only the

frames in which both properties are easy to judge. Motion intervals and difficult transitions are part of the task rather than inconvenient observations to remove.

A useful protocol would include multiple expressions and repeated trajectories under declared initial conditions. It would preserve the order and duration of the requested changes. A single selected smile does not establish control over a continuing sequence, and a static average over frames can conceal a brief substitution. The analysis should specify whether a local error invalidates a clip, contributes to a temporal rate, or is reported through another prespecified rule. That choice determines what a final preservation score represents.

The choice of statistical unit should precede the choice of a standard error. A dataset can contain many rows but relatively few independently acquired histories. Heteroskedasticity-consistent covariance protects an otherwise specified regression against unequal residual variances (White, 1980). It does not automatically accommodate observations sharing an environment, subject, or recording. Cluster-robust inference makes that dependence part of the variance calculation (Cameron and Miller, 2015). The grouping must follow the observation process and the claim being made; an identifier that is convenient to store is not necessarily the relevant independent unit. For a descriptive analysis, preserving the groups also allows readers to recalculate a different population average. An uncertainty interval remains conditional on the observations and assumptions actually available.

Discriminative identity representations address a different question from the design of the test population. FaceNet learns an embedding for verification and related recognition tasks (Schroff et al., 2015). ArcFace (Deng et al., 2018) and CosFace (Wang et al., 2018) develop margin-based approaches to discriminative face learning. These methods do not determine which reference views should be supplied to a generator. Dataset choices also alter the challenge: VGGFace2 emphasizes pose and age variation (Cao et al., 2017), MegaFace increases the scale of recognition distractors (Kemelmacher-Shlizerman et al., 2015), and FairFace studies representation across annotated demographic groups (Kärkkäinen and Joo, 2019). Their relevance is to separate distinct kinds of coverage. Many comparisons of a small reference set cannot replace missing pose variation or a different population of people. A study should state which identity measurement it uses, how thresholds are selected, and what the represented cases actually cover. Strong performance on one protocol does not establish universal validity of the embedding across generated or stylized subjects.

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. Measurement sensitivity and valid ambiguity

An automated identity instrument may respond to expression changes even when human observers recognize the same person. That possibility needs direct measurement rather than an assumption that the instrument is wrong or that the generator has drifted. A fixed set of outputs could be assessed by several declared procedures, with disagreements preserved for independent interpretation. The relevant population is the expression and motion range actually included in the comparison.

Human observers also need suitable evidence. If they see only a heavily occluded transition frame, they may reasonably decline a confident identity judgment. That should be distinguished from a verdict that the person changed. Coverage can therefore vary across the trajectory. A report should preserve when judgments are available and examine whether a method appears strong mainly because difficult frames are unscorable. The same caution applies to automatic detection failures before an embedding is calculated.

The proposed design should avoid treating expression stability as the only desirable outcome. Some applications require a specific expressive instruction; others permit a range of valid performances. The task specification should state that range and the consequences of failing it. A generator that produces a recognizable subject with little requested expression may be useful in one workflow and inadequate in another. Evaluation should make that tradeoff visible rather than assume that maximum similarity to a neutral reference is always the correct objective.

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 proposed explanation should be distinguishable from alternative explanations using observations the study can actually collect. Diagramming the relationships among selection, exposure, measurement, and outcome makes that requirement explicit (Greenland et al., 1999). The diagram does not establish its own correctness, but it prevents adjustment choices from remaining invisible. The same transparency is needed when investigators inspect many endpoints before choosing a narrative. Controlling false discovery concerns the behavior of a specified testing family (Benjamini and Hochberg, 1995); it is not a general certificate of scientific validity. Primary endpoints, secondary descriptions, and future mechanism tests should be named separately. Readers can then assess whether an attractive pattern was predicted, explored, or merely noticed after the available comparisons were examined.

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

An identity evaluation becomes more informative when it retains the other properties of a generated clip. VBench's multiple dimensions (Huang et al., 2023) help motivate this decomposition. Structural similarity (Wang et al., 2004) compares a particular aspect of image structure; learned perceptual similarity (Zhang et al., 2018) uses another representation; video-distribution metrics (Unterthiner et al., 2018) incorporate temporal visual features. None alone establishes a stable person through every pose and event. A study should identify the relevant time windows and examine local failures as well as sequence averages. It should also distinguish an identity evaluator's inability to score a frame from the generator's inability to preserve a subject. Those are different observations with different remedies. A measurement profile linked to the original clip permits this distinction to be checked and avoids attributing broad capability to a favorable average obtained only on easily measurable views.

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.

An evaluation remains comparable only if its inputs and transformations can be identified. The FAIR principles emphasize persistent identification and reusable scientific metadata (Wilkinson et al., 2016). Work on machine-learning technical debt highlights how dependencies and hidden couplings accumulate in a pipeline (Sculley et al., 2015). In this setting, a score should be linked to the exact references, generation settings, evaluator version, and aggregation rule that produced it. A replacement checkpoint or a revised eligibility filter may change the scientific question even when the output column keeps the same name. Preserving such changes in a release record permits later comparisons to be interpreted honestly. The objective is not maximum documentation volume, but a recoverable account of the decisions that affect the result.

5. Conclusion

Expression tests should ask whether the subject is preserved through the change the task requires. They need separate observations of expression execution, identity, temporal behavior, and measurement coverage. The proposed design also records which expressive evidence a model received in its references. It does not establish a new result for ARGUS or another generator. Its value is to prevent a preservation score from rewarding an unperformed transformation or confusing a legitimate expression change with an unverified identity substitution.

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.