Missing Views and Honest Coverage in Subject-Preserving Generation

Christopher Butler (MD), Tyler Young (VE), Oliver Barrett (PT)

Abstract

A reference set cannot provide every possible view of a subject. An evaluation should distinguish performance within its observed views from behavior when requested evidence is missing. This article proposes a coverage record and a withheld-view study for reference-conditioned generation. ARGUS supplies the motivating multi-view setting. No views are removed and no clips are generated here. The design makes missing evidence visible without assuming that every unsupported view is either a demonstrated identity failure or a successful generalization.

1. Reference absence is part of the task

A subject may be represented by several frontal images and no useful profile, or by one session with limited expression variation. Counting references alone can obscure these gaps. A generation instruction that moves beyond the supplied observations asks a different question from one that reproduces a familiar view. A benchmark should describe that difference before averaging the results, since the apparent preservation capability depends on the relationship between the requested output and available reference evidence.

ARGUS structures multiple identity views for video generation (Meng et al., 2026b). That design motivates an examination of reference coverage rather than an assumption that a multi-view label implies complete support. The present article proposes a missing-view study and reports no generation outcomes. It asks how to distinguish an absent input observation, an unassessable output, and an independently observed failure of subject preservation when a requested view lies outside the supplied reference set.

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.

The validity of an identity evaluation depends partly on how its scientific objects are named and linked. LFW's identity, image, pair, and fold units are explicit parts of its protocol (Huang et al., 2008). The scikit-learn implementation documents which file corresponds to each supported task (Pedregosa et al., 2011). Dataset provenance should similarly make the intended use and selection process recoverable (Gebru et al., 2021). In video generation, an evaluator needs additional links among reference observations, subjects, clips, and frames. Repeated references can provide useful evidence without becoming independent subjects. A pair-level metadata audit can reveal those relationships, but it cannot evaluate the visual content of a generated clip. The distinction is productive: data organization can be checked before an expensive model study begins, and the missing measurements can be specified rather than silently assumed to exist.

2. A withheld-view comparison with a defined boundary

A proposed collection would define view categories or measured poses using a documented procedure. The reference inventory would retain uncertain assignments rather than force every image into a precise category. A controlled comparison could withhold a declared view while retaining matched source information as far as feasible. The full and reduced sets would be recorded for each subject so that the actual change in evidence can be reconstructed.

Withholding a view often changes more than geometric coverage. It may remove the highest-resolution image, a distinctive expression, or a unique lighting condition. A useful design would document these simultaneous changes and either control them through collection or acknowledge them in interpretation. An adjusted observational analysis cannot manufacture the missing combination. Its evidence would concern the particular reduction in reference information rather than an isolated causal effect of viewpoint unless the design supports that stronger attribution.

Assessment requires independent evidence about the intended subject in the requested view. If evaluators only see the reduced references, they may share the generator's uncertainty. A separate assessment condition could supply held-out views, but that changes the information available to the assessor and should be explicit. Both conditions can be informative: one describes the user-facing evidence boundary, while the other examines identity using additional observations unavailable during generation.

An association can suggest a research direction without identifying an intervention. Causal diagrams make it possible to state which variables precede the proposed change, which are consequences, and which selection processes remain uncontrolled (Greenland et al., 1999). Their value lies in exposing assumptions rather than decorating a regression with a causal label. A second issue arises when many outcomes or subgroups are explored: selecting the most favorable comparison can exaggerate the impression of support. False-discovery procedures address a defined family of inferential tests (Benjamini and Hochberg, 1995). They do not turn an exploratory pattern into a replicated mechanism. A careful report distinguishes prespecified comparisons, descriptive sensitivities, and hypotheses requiring new observations. This distinction is especially important when an evaluation offers many plausible metrics.

A face embedding is an instrument whose output needs a defined task. FaceNet supports verification and recognition through a learned metric space (Schroff et al., 2015), while ArcFace (Deng et al., 2018) and CosFace (Wang et al., 2018) pursue discriminative margins through different formulations. The scientific population is supplied by the benchmark, not by those objectives. VGGFace2's within-person variation (Cao et al., 2017), MegaFace's distractor scale (Kemelmacher-Shlizerman et al., 2015), and FairFace's representation emphasis (Kärkkäinen and Joo, 2019) show why populations can differ materially. For subject-preserving video, reference access and output conditions should therefore be documented alongside the identity score. An automated evaluator may be accurate on its original photographic population yet behave differently on generated motion, unusual lighting, or stylization. Establishing the relevant transfer requires observations in those settings. Architectural sophistication in the identity instrument cannot replace a direct assessment of the claim the generator is making.

3. Reporting coverage and generalization together

The primary report would link outcomes to the reference-output relationship. Results within well-supported views should remain distinguishable from extrapolative requests. A single pooled mean can conceal a method that performs well only where references are abundant. Conversely, a poor automated score on unsupported views may reflect an inadequate evaluator. Coverage records and independent judgments help separate these possibilities without assuming that missing observations establish the direction of the result.

The comparison should retain generation failures, unavailable identity scores, and uncertain human judgments. A method that produces fewer assessable outputs can otherwise appear strong on its retained subset. An explicit missing-measurement policy makes that selection visible. Sensitivity summaries can illustrate alternative treatments of uncertainty, but they should not assign observed truth to cases where evidence remains absent. The unresolved portion belongs in the final statement of what the study supports.

Repeated samples and reference variants should remain grouped by subject and task. Multiple generated views of one person do not establish broad population coverage, even if their frame count is large. The study should identify which sources of variation it samples and which are fixed. A useful conclusion can then be bounded to the documented views, subjects, and reference conditions, while future collections target the gaps that the coverage audit has made explicit rather than merely increasing the number of similar images.

The reference path and the scoring path should be distinguishable in an experiment. DreamBooth (Ruiz et al., 2022) and IP-Adapter (Ye et al., 2023) illustrate different ways to make subject or image evidence influence generation. CLIP (Radford et al., 2021) and DINO (Caron et al., 2021) illustrate representations that can be used to compare visual content for other purposes. Reusing the same representation for conditioning and evaluation may create shared sensitivities that deserve examination. It is not proof of invalidity, but it strengthens the case for an independent human or task endpoint. A comparison should report which information enters each path and retain reference-specific artifacts as potential confounds. The generator's ability to reproduce a background, camera style, or distinctive clothing can be useful while remaining different from preserving identity under a requested change. A careful evaluation makes that difference observable.

An evaluation protocol should state what remains unfamiliar at assessment time. Leaving out an observation while retaining its close relatives can test interpolation rather than the broader transfer suggested by the study's wording. Validation methods for structured data emphasize matching the split to temporal, spatial, or hierarchical dependence (Roberts et al., 2017). Separately, parameter selection can bias an evaluation if the same observations guide both choices and reporting (Varma and Simon, 2006). These issues are not solved by calling a split a test set. The record should identify shared sources, the information accessible during development, and the stage at which each design choice was fixed. A result can be informative within a narrow boundary while providing no direct evidence about a more distant deployment population.

Coverage should be reported whenever the system or its evaluator can avoid difficult cases. An abstention policy changes the population on which an accuracy or fidelity score is calculated. Selective classification formalizes the joint reporting of coverage and conditional performance (Geifman and El-Yaniv, 2017). Distribution-shift research adds that estimated uncertainty can behave differently outside the development population (Ovadia et al., 2019). A benchmark should therefore retain which observations were eligible, accepted, rejected, or unavailable. A result on a carefully limited subset can be useful, provided the subset

is visible. The challenge is to prevent a narrow success claim from becoming a broad capability claim through omission of the denominator or the circumstances in which the system declines to operate.

4. Connecting an identity claim to its observations

A moving subject introduces quality dimensions that a still-image identity score cannot exhaust. VBench decomposes video-generation evaluation into multiple aspects (Huang et al., 2023). Structural similarity (Wang et al., 2004), learned perceptual correspondence (Zhang et al., 2018), and video-distribution metrics (Unterthiner et al., 2018) provide additional observations under their respective assumptions. They should not be treated as equivalent measurements of the same latent property. A generated clip can have coherent motion and attractive appearance while changing its subject; it can also preserve recognizable identity while containing visible temporal artifacts. Reporting separate outcomes allows those tradeoffs to be examined. The assessment should specify the frames, viewpoints, occlusion states, and durations included in each score. A difficult case that defeats the identity evaluator should remain part of the coverage record rather than disappearing silently. The resulting profile can support a bounded claim about the observed conditions without implying that every aspect of subject consistency has been solved.

An automated measurement should not inherit the authority of its training labels without a direct assessment of use. The relevant checks include systematic disagreement with the intended procedure (Bland and Altman, 1986) and calibration of any probability claims (Guo et al., 2017). Ranking performance, absolute accuracy, and confidence reliability are distinct outcomes. A study can improve one while leaving the others unchanged. Keeping the measurements separate permits examination of those tradeoffs and avoids implying that one favorable statistic establishes all aspects of validity. When independent judgments are unavailable, the limitation should remain part of the result. The absence of a target observation cannot be repaired by combining several correlated proxies into a more elaborate total score.

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

Missing views should be represented as a property of reference evidence and assessment coverage. The proposed design distinguishes withheld inputs, extra evaluator information, uncertain outputs, and observed identity failures. It offers no new generalization result. Its contribution is a transparent way to investigate what a generator does beyond supplied views while keeping unsupported conditions visible, allowing limited successes to be understood without extending them to every unobserved appearance of a subject.

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.