Reference Overlap and Contamination in Identity Benchmarks

Calvin Rhodes (KH), Aaron Roberts (KR), Benjamin King (ME)

Abstract

Shared references across benchmark partitions can be detected from records, whereas overlap with a model's training data may remain partly unknown. Conflating the two produces either unwarranted confidence or an overly broad contamination claim. This article develops a layered audit for subject-preserving video evaluation, motivated by ARGUS. No training corpus is inspected. The proposed audit distinguishes exact artifact reuse, related source observations, subject overlap, and unverified training exposure, assigning each conclusion to the evidence that could actually support it.

Overlap has more than one unit

Two benchmark entries can share an identical image, different crops of one observation, different images of the same person, or only a related acquisition context. These relationships matter in different ways. Subject overlap may be intentional in a task that evaluates new views of known people, while exact source reuse can undermine a claim about unseen observations. An audit must define its intended separation before treating every shared identity or filename as the same problem.

ARGUS organizes identity references for subject-preserving generation (Meng et al., 2026b). That setting motivates an overlap audit connecting references, generated tasks, and assessment records. This article proposes the audit without inspecting training data or claiming a contamination finding for the focal model. Its question is which overlap relationships can be established from documented artifacts and how an evaluation should state the remaining uncertainty when the model's complete training history is unavailable.

A layered audit with explicit evidence

A proposed first layer would compare exact file identities using hashes and stable identifiers. This can establish identical artifact reuse when the artifacts are available, but it will miss changed encodings or crops. A second layer could examine documented parent observations and source sessions. Visual duplicate detection may help identify candidates, yet its matches require a declared review process and should retain uncertainty rather than automatically becoming verified source links.

A third layer would examine subject overlap across development and assessment partitions. The intended benchmark task determines its meaning. A new-subject test needs a different boundary from a new-view test of familiar subjects. The audit should report both the observed overlap and the declared task requirement. It should avoid presenting a clean image split as a clean subject split or implying that a subject-disjoint assessment automatically excludes every related visual observation from all development decisions.

Training exposure requires another evidence source. A public image's availability does not establish that a particular model encountered it, and failure to find it in an incomplete corpus does not prove absence. Known training records, documented collection rules, or appropriate direct checks can support bounded conclusions. If those records are unavailable, the audit should identify that boundary. It can still establish benchmark-internal overlap without converting that useful result into a claim about the model's unseen training history.

How overlap changes an evaluation claim

A detected relationship should be connected to the specific generalization statement it threatens. Exact reference reuse may be compatible with a task designed to generate new motion from known images, but it does not establish performance on unseen source observations. The report should describe the narrower task directly. Removing related entries can support another comparison, provided the resulting population and sample size are recorded rather than silently replacing the original evaluation denominator.

Overlap checks should occur before outcome-driven selection where feasible. If suspicious entries are removed only after poor results are visible, the audit can become part of benchmark tuning. A transparent procedure would preserve the rule, candidate list, adjudication, and version change. Sensitivity analyses on known overlap categories can be informative, but their interpretation should remain conditional on those categories and should not imply that every undetected relationship has been ruled out by the search.

Finally, records should distinguish benchmark integrity from generated performance. A clean separation supports a particular evaluation boundary; it does not establish high identity accuracy. A contaminated or uncertain boundary can limit a broad claim while leaving some descriptive output observations useful. This layered interpretation avoids an all-or-nothing verdict and directs follow-up work toward the missing evidence, whether that is source provenance, subject linkage, a revised split, or a more complete training record.

Checking overlap at more than one level

For the proposed comparison, the key observation is source and partition identity. A training partition contains a differently cropped portrait from the same session as a benchmark reference. This hypothetical case poses a question about which assessment independence claims the available links substantiate. Record exact-file duplication check and source-session and subject-linkage audit separately. A different hash can conceal visual or source overlap while a shared subject may be intended in a new-view task. Report exact artifact, related source and verified training exposure separately.

Use exact hashes to establish identical files and retain acquisition-session or derivative links for broader overlap. A benchmark for a new pose may permit the subject to occur in development while holding out capture views. A new-subject claim has a stricter boundary. Record those relations and evaluate each split against its stated objective. Training exposure can be unknown even after the benchmark files are shown disjoint. Reserve a separate category for unverifiable exposure rather than inferring it from the public availability of the photographs.

Before estimating which assessment independence claims the available links substantiate, fix the conditions under which a case counts as eligible and decide how exclusions will appear in the summary. For the source and partition identity, keep undetectable faces in the log and state which cases remained eligible for recognition.

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

Use the comparison with exact-file duplication check to identify which parts of the finding survive a reasonable change in the decision threshold. The evaluation of exact-file duplication check can distinguish an incorrect subject from a permitted change in expression, pose or scene. Document the sampling

boundary for which assessment independence claims the available links substantiate and reserve a new cohort of cases for a distinct replication.

Samples and evidence

Validation should respect shared identities, scenes or capture times when units near one another would otherwise leak into nominally held-out data. Pose and age diversity help construct cross-condition verification pairs, but transfer tests still need outputs from the intended workflow. An existing method connects validation splits to temporal, spatial and hierarchical dependence (Roberts et al., 2017). For a prospective comparison, an existing method provides face observations with substantial variation in pose and age (Cao et al., 2017).

Measurement and error

Related work combines detection and association considerations in tracking assessment (Luiten et al., 2020). Earlier research investigates scaling in contrastive language-image learning (Cherti et al., 2022). Report the detection and association components as well as the aggregate so that tracking continuity can be interpreted alongside object availability. In practical terms, representation changes can alter relative prompt scores, so the encoder checkpoint and text preprocessing belong in the comparison record.

Low prevalence can make a seemingly small false-positive rate operationally costly, so report precision at plausible recall levels and retain class counts. Its separate video dimensions provide a reason to publish a failure profile instead of combining every observation into a single quality index. For implementation, another study organizes video-generation evaluation into distinct measurement dimensions (Huang et al., 2023). An existing method contrasts precision-recall and receiver operating characteristic summaries under class imbalance (Saito and Rehmsmeier, 2015).

Related work learns visual representations through self-supervision (Caron et al., 2021). Another study learns aligned visual and language representations through contrastive training (Radford et al., 2021). In practical terms, a whole-image embedding reflects scene description and background alongside subject appearance. A local check can distinguish these contributions. Using a representation learned without explicit class labels gives a useful comparison when a prompt-derived score may reward the wrong visual feature.

A comparison of video systems should record differences in training material and processing before attributing a visible change to the architecture alone. One published study uses a cosine margin for discriminative face recognition (Wang et al., 2018). For implementation, a cosine-margin baseline helps distinguish rejection due to the representation from rejection caused by the threshold. A methodological precedent investigates data and training choices for text-to-video synthesis (Chen et al., 2024).

Angular discrimination makes a reasonable verification baseline, provided the threshold is evaluated again on profiles, occlusions and synthetic views outside ordinary portrait capture. Object annotations support localized checks, with the class mapping and annotation policy recorded before evaluating cases absent from the detector vocabulary. In a controlled test, one study uses an additive angular margin to train discriminative face embeddings (Deng et al., 2018). Earlier research provides object annotations and standardized visual recognition tasks (Lin et al., 2014).

A general visual representation can provide an independent similarity score whose errors may differ from those of a face or prompt-specific model. Landmark localization provides a way to identify crop failures before measuring resemblance and to retain a denominator for faces that were never detected. Another

study jointly localizes faces and facial landmarks (Deng et al., 2019). In an evaluation, a methodological precedent learns general-purpose visual features from curated image data (Oquab et al., 2023).

Systems and comparators

Image-conditioned video provides a direct persistence test. Keep the supplied view and later output segment as separate observations. A relevant contribution uses appearance information in online association (Wojke et al., 2017). An existing method investigates image-to-video generation and video training data curation (Blattmann et al., 2023). During validation, an appearance-aware baseline can be compared with motion-only association under the same detections to locate the source of identity changes.

A one-reference adapter supplies a baseline for evaluating whether added views help temporal persistence when motion and instructions are held constant. A learned concept embedding offers a compact control for comparing changes in prompt adherence under the same set of reference images. One approach personalizes video through a face adapter using one reference (He et al., 2024). For implementation, another study learns concept embeddings from a few images (Gal et al., 2022).

Analysis and interpretation

Single-reference conditioning provides a budget-matched comparison for identifying which errors disappear when extra views are made available to the generator. One study separates missing-data assumptions from observed-data summaries (Little and Rubin, 2002). Related work conditions identity-preserving image synthesis on a single reference (Wang et al., 2024). In practical terms, a missingness analysis distinguishes failures from observations that were never eligible and makes the assumptions behind complete-case summaries visible to readers.

Earlier research reviews matching as a design step for comparing observational units (Stuart, 2010). Related work shows bias from reusing validation for selection and error estimation (Varma and Simon, 2006). In practical terms, matching can improve comparability on recorded attributes; an overlap display should identify which observations are lost before the outcome contrast is computed. Reusing a validation split for repeated configuration selection can make the final error estimate optimistic; retain a separate assessment for the chosen rule.

Linking evidence to the evaluation

Pair definitions determine the verification denominator, so any comparison should state how multiple views of the same person were sampled. During validation, an existing collection provides a starting point for auditing subgroup coverage while leaving the deployment distribution to be specified separately. An existing method constructs a face attribute dataset with demographic balance (Kärkkäinen and Joo, 2019). A relevant contribution defines an unconstrained face verification collection and pair protocol (Huang et al., 2008). A relevant contribution builds a stacked identity representation from input portraits (Li et al., 2023a). One published study uses causal diagrams to articulate assumptions about adjustment and selection (Greenland et al., 1999).

Earlier research tests face recognition with a large distractor collection (Kemelmacher-Shlizerman et al., 2015). Another study specifies principles for discoverable and reusable research artifacts (Wilkinson et al., 2016). One study asks dataset creators to record collection, intended use and limitations (Gebru et al., 2021). In an evaluation, a reproducible result requires identifiers for both the released inputs and the exact intermediate tables from which the paper derives its figures. The size and makeup of the distractor gallery alter the false-match burden, so gallery construction belongs in the main experimental description.

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

Conclusion

Overlap audits need explicit units and separate evidence for artifacts, source observations, subjects, and training exposure. The proposed design reports no new contamination finding. It allows benchmark-internal checks to produce useful bounded conclusions while preserving uncertainty about unavailable training history. A subject-preservation result should then state precisely which separation its records establish and which broader generalization claims remain unsupported by the audit performed.

References

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.

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

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.

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.

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.

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.

Oquab, Maxime, et al. "DINOv2: Learning Robust Visual Features without Supervision." *arXiv preprint arXiv:2304.07193*, 2023. https://arxiv.org/abs/2304.07193.

Blattmann, Andreas, et al. "Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets." *arXiv preprint arXiv:2311.15127*, 2023. https://arxiv.org/abs/2311.15127.

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.

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.

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.

Cherti, Mehdi, et al. "Reproducible scaling laws for contrastive language-image learning." *arXiv preprint arXiv:2212.07143*, 2022. https://arxiv.org/abs/2212.07143.

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.

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

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.

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.

Chen, Haoxin, et al. "VideoCrafter2: Overcoming Data Limitations for High-Quality Video Diffusion Models." *arXiv preprint arXiv:2401.09047*, 2024. https://arxiv.org/abs/2401.09047.

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

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.

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.

Huang, Ziqi, et al.. "VBench: Comprehensive Benchmark Suite for Video Generative Models" *arXiv preprint arXiv:2311.17982*, 2023. https://arxiv.org/abs/2311.17982.

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.

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.

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.

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.

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