Binding Multiple Subjects to Their References across a Scene

Benjamin Walsh (DJ), Kenneth Rhodes (MX), Ian Peterson (FR)

Abstract

A video can preserve several recognizable faces while assigning them to the wrong roles or trajectories. This conceptual study proposes an explicit mapping between subject references, scene roles, and time-indexed outputs. ARGUS motivates the identity-conditioning mechanism; the article does not assume that its reported single-subject evidence establishes multi-subject binding. No videos are generated here. The proposed evaluation separates individual recognizability, assignment correctness, and temporal continuity, making reference swaps and duplicated subjects visible instead of allowing a high average face score to stand in for correct scene composition.

1. Recognizable parts can form the wrong scene

Suppose two referenced subjects are requested in different roles. An output may contain both recognizable people while swapping their actions, positions, or dialogue roles. A pooled similarity score over all detected faces can remain favorable because the right ingredients are present. The task nevertheless fails if the reference-to-role assignment is required. Identity preservation within a scene therefore includes a binding question that cannot always be answered by assessing each face independently.

ARGUS provides a motivating reference-conditioning approach (Meng et al., 2026b). This article extends the evaluation question conceptually to scenes containing multiple subjects, without claiming that the paper has already established such support. It performs no multi-subject generation or assignment analysis. The contribution is a task record and proposed endpoint structure through which a future study could distinguish recognizable individuals from correctly assigned, persistent scene participants.

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.

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.

2. The mapping that precedes a binding score

A proposed task would identify each reference set and associate it with a declared scene role. The role might concern a spatial relation, an action, or another observable requirement. The prompt should make that mapping unambiguous to an independent evaluator. If the instruction itself leaves the role open, several assignments may be valid. The study should not choose a preferred interpretation only after inspecting the output. A supported ambiguity rule is part of the task contract.

The output record would preserve time-indexed subject observations and their proposed associations. Individual recognizability asks whether the intended subject is present in an observation. Binding asks whether that observation occupies the required role. Continuity asks whether the association persists over the relevant events. A duplicated subject can satisfy one reference match several times while leaving another role unfilled. A swap can preserve the set of identities while violating continuity. The report needs endpoints that retain these distinctions.

Reference access also needs to be documented per subject. One person may have several high-quality views while another has only a difficult observation. A system may allocate memory, generation effort, or manual selection unequally. Those differences can affect both identity and assignment outcomes. A fair comparison should match the available information where feasible or state the workflow advantage directly. Equal numbers of named subjects do not establish equal evidence for each one.

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.

The recognition literature provides both scoring methods and warnings about incomplete evaluation populations. FaceNet's learned embedding (Schroff et al., 2015), ArcFace's angular-margin formulation (Deng et al., 2018), and CosFace's margin-based discrimination (Wang et al., 2018) are alternative measurement foundations rather than interchangeable proof of personal identity. VGGFace2 introduces substantial pose and age variation (Cao et al., 2017), MegaFace studies recognition at large scale (Kemelmacher-Shlizerman et al., 2015), and FairFace emphasizes representation in annotated face datasets (Kärkkäinen and Joo, 2019). An identity-preserving generator should be assessed with the relevant dimensions recorded separately. A result from familiar frontal images cannot establish reliability under a missing viewpoint. Nor can a frequency-balanced set establish demographic balance without the required labels. The evaluation should retain failures of detection or measurement, identify what threshold was used, and explain whether the endpoint concerns verification, identification among distractors, or human recognizability over a moving sequence.

3. Following assignments through interaction

The proposed evaluation should include events that exercise the mapping. Subjects can cross, become occluded, exchange positions, or perform similar actions. These events should be designed around a declared question rather than gathered only from visually impressive demonstrations. A trajectory-level assessment can then identify whether a role follows the intended person or merely follows a location in the image. A correct static arrangement at the beginning would not establish the required continuity.

Automated assignment needs an independent validation boundary. If the same representation supplies both identity similarity and tracking association, shared errors can produce internally consistent but incorrect labels. A study should retain ambiguous matches and obtain independent judgments where the task requires them. It should also distinguish an absent subject, an unmeasurable subject, and a confidently wrong association. Forcing every detected face into one of the requested references can hide the possibility that the generator introduced a different person.

Coverage should be reported at both subject and scene levels. A method may generate one role well and repeatedly fail another, while an average obscures the imbalance. A scene-level success rule can make joint requirements visible, but its relationship to per-subject outcomes should remain explicit. The appropriate weighting depends on the intended workflow. A scene requiring every role to be correct has a different endpoint from a collection used to select one acceptable subject crop afterward.

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 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.

A system's reliable operating region may be narrower than the collection of inputs an interface accepts. Selective prediction separates performance on accepted cases from the fraction of cases accepted (Geifman and El-Yaniv, 2017). Reporting only the former can hide a substantial loss of coverage. Uncertainty estimates may also change under distribution shift (Ovadia et al., 2019), so familiar-condition confidence is not automatically evidence about new subjects, scenes, or control histories. A useful evaluation records rejected, ambiguous, and unsupported cases alongside successful ones. This makes it possible to distinguish limited but honest capability from a system that answers broadly with unreliable

certainty. Any proposed deferral rule still requires assessment on its intended input population; it is not validated merely by being described.

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.

Replacing an expensive judgment with an automated score changes the measurement process. A positive correlation is evidence of co-movement, but it does not establish that the two procedures can be used interchangeably (Bland and Altman, 1986). A calibrated forecast has a different requirement: its stated probabilities should match relevant observed frequencies (Guo et al., 2017). Neither property follows merely from a favorable overall ranking. A useful comparison preserves paired measurements, identifies the conditions in which disagreement matters, and keeps scale transformations explicit. If the intended judgment concerns recognizability, playability, or physical compatibility, the proxy should be assessed against that judgment directly. Agreement on easy cases cannot supply evidence about the difficult conditions absent from the comparison.

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

Multi-subject preservation needs a recoverable mapping among references, roles, and trajectories. The proposed design separates recognizability from binding and continuity, retains missing and ambiguous observations, and documents evidence access per subject. It does not report a new ARGUS capability. Its contribution is to ensure that a future scene-level result concerns the required composition of people rather than only a favorable set of independent reference similarities.

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.