Conflicting References in Text-Conditioned Subject Generation

Owen Evans (IQ), Ryan Hall (PL), Grant Barker (BE)

Abstract

A prompt, first frame, and identity reference can request incompatible appearances. This conceptual study proposes an explicit precedence contract for those conditions and an evaluation that distinguishes valid resolution from failed instruction following. ARGUS motivates the reference-conflict problem. No conflicting inputs are submitted to a model here. The design preserves which requirement is authoritative, which changes are allowed, and whether the resulting output can be judged. It argues that a plausible compromise is not automatically a correct solution when the interface never specified what the competing conditions were supposed to mean.

1. An inconsistency can belong to the request

A reference may show one accessory while a prompt requests another. A first frame can establish a scene that conflicts with the text description. Some inconsistencies are intentional editing tasks; others are accidental or outside the supported input contract. A generator cannot be evaluated fairly unless the protocol states which interpretation applies. Calling every compromise successful would make instruction following untestable, while calling every unresolved inconsistency a generator error would ignore invalid requests.

ARGUS describes identity conditioning that includes selection and resolution of competing information (Meng et al., 2026b). That makes precedence a useful research question, but does not give this article a verified contract for every possible input combination. The present work is a conceptual evaluation design. It does not run the framework or reinterpret an observed conflict outcome. Its aim is to define the task before scoring the output, so the meaning of successful identity preservation remains compatible with the other instructions.

A reference collection can grow in several ways: more identities, more images of existing identities, or more comparisons among the same images. The LFW protocol makes these units distinguishable (Huang et al., 2008), while its official loader documentation preserves the difference among development and assessment files (Pedregosa et al., 2011). Datasheets provide a framework for explaining why those objects were collected and which uses they support (Gebru et al., 2021). A subject-generation study should apply the same discipline to reference mosaics and output trajectories. The number of reference tiles is not automatically the number of independent observations, and the number of scored frames is not automatically the number of subjects. Recording the hierarchy permits appropriate weighting and uncertainty calculations later. Where a study uses only a public pair list, its conclusions should remain about that list and its protocol. They cannot supply observations of appearance, motion, or recognition errors that were never processed.

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.

2. The contract that an evaluator needs

A proposed test would label each input condition by its role. Identity evidence might be authoritative for the subject, while scene text specifies a change to setting. A first frame could be a required initial observation or a loose visual suggestion. Those are different tasks. The protocol should identify which aspects of each source are fixed, editable, or optional. A requested change in clothing should not be mistaken for permission to change the person, unless that is explicitly the task.

The test collection would include compatible cases, supported conflicts with a declared resolution rule, and unsupported combinations. Their expected outcomes can differ: execution, clarification, or rejection may each be appropriate. The report should preserve acceptance and rejection decisions. A system that narrows its input support can improve conditional output quality without demonstrating broader control. Conversely, rejecting an invalid request should not automatically count as a failure to preserve identity in a task the interface never accepted.

Scoring should separate identity, requested transformation, and precedence compliance. An output can retain the person while ignoring the authoritative scene instruction. It can satisfy text while copying a different identity from the first frame. A single generic similarity score may reward whichever source its representation happens to emphasize. Independent endpoints and a recoverable precedence rule make those outcomes interpretable. The study should also retain ambiguous outputs rather than force a verdict that the evidence does not support.

Replication has several levels: a second measurement, a second trajectory, and a second independently acquired setting are different additions to evidence. Statistical calculations should retain those distinctions. Heteroskedasticity-robust regression inference concerns unequal conditional variances (White, 1980); cluster-robust inference concerns dependencies shared within groups (Cameron and Miller, 2015). The group definition should be justified through the data collection and evaluation process, and not chosen because it produces a preferred significance result. Unequal group sizes also influence what a pooled average represents. Reporting both the group count and observation count permits a reader to see whether the calculation describes common cases, heavily recorded cases, or an explicitly balanced group population. No variance estimator alone resolves that choice of estimand.

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.

Subject conditioning is not identical to measuring the resulting subject. DreamBooth studies subject-driven generation through model adaptation (Ruiz et al., 2022), whereas IP-Adapter develops image prompting compatible with text conditioning (Ye et al., 2023). The reference evidence available to those mechanisms should be matched or explicitly described in a comparison. Generic representations

also have different training purposes: CLIP connects imagery to language (Radford et al., 2021), while DINO investigates self-supervised visual features (Caron et al., 2021). Either can contribute to a measurement pipeline, but neither automatically defines human identity across generated motion. A study should separate input fidelity, recognizability, and obedience to scene instructions. It should also retain cases in which these objectives conflict. An image-conditioning mechanism can reproduce an accessory or background very well while leaving a more important identity requirement unresolved. Measuring those conflicts directly is more informative than declaring every increase in reference similarity a successful preservation of the person.

3. Testing resolution without changing every condition

A diagnostic design could vary one conflicting source while holding the other information fixed, with repeated outputs under declared sampling conditions. It would measure whether the authoritative requirement remains stable and whether the permitted change occurs. This does not require exact pixel identity across generations. It requires a task rule that distinguishes correct and incorrect consequences. A visible difference alone would not establish meaningful resolution of the intended conflict.

Selection and generation should also be examined separately where possible. An internal controller may select the right reference but the resulting clip may still violate the contract. An incorrect internal explanation can also accompany an externally acceptable output. The evaluation should preserve both observations without treating agreement among model components as independent confirmation. The external task consequence remains the principal evidence about the completed workflow.

The scope of the result should be tied to the tested conflict types. A system can support changes in accessories while remaining untested on conflicts involving viewpoint, scene geometry, or multiple subjects. The record should identify this support instead of translating success on a few curated examples into universal instruction arbitration. A useful benchmark makes those boundaries visible and supplies a structured path for testing further conditions in new cases.

The phrase generalization needs an object: another observation of the same source, another source within the same collection, or a different acquisition process. Each requires a different assessment boundary. Structured validation makes explicit how correlated observations are assigned across training and evaluation (Roberts et al., 2017). Selection bias can still arise if the final boundary is repeatedly consulted while choosing a method (Varma and Simon, 2006). Publishing the split identifiers and the information permitted during each stage allows the claimed transfer to be checked. When a study proposes a new architecture but performs only a data audit, it should state that difference directly. Useful evidence about data coverage does not require pretending that a prediction system has already been trained and externally validated.

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.

4. Reference access and measurement reliability

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.

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.

Reproducibility begins with preserving which scientific objects are being linked. Stable identifiers and reusable metadata help connect inputs, reference observations, model versions, and reported outcomes (Wilkinson et al., 2016). The scoring pipeline also has dependencies that can change the interpretation of a result: preprocessing, representation checkpoints, thresholds, and aggregation code can all matter (Sculley et al., 2015). A compact release should retain those decisions rather than only the final mean. Where source data cannot be redistributed, an access description and hashes can still identify the analyzed inputs. This record allows later users to distinguish a change in the underlying system from a change in how the system was measured. Reproducibility is therefore part of interpretation, not simply a final packaging step.

5. Conclusion

Reference conflicts need a task contract before they need a winning output. The proposed design distinguishes authoritative identity evidence, editable appearance, valid combinations, and unsupported requests. It then measures resolution, preservation, and coverage separately. No new model result is claimed. The contribution is an interpretable evaluation boundary through which a subject-generation workflow can demonstrate the particular kinds of conflicting conditions it actually handles.

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.