Preserving Expression as well as the Subject in Generated Motion

Robert Stevens (AZ), Kyle Turner (AO), Connor Palmer (AO)

Abstract

A subject can remain recognizable while a requested expression fails, and an expression can be reproduced while the person changes. This article develops a study design that treats these as separate outcomes. ARGUS motivates reference-conditioned motion with identity continuity. No expressions are annotated and no videos are generated here. The proposed comparison records the expression trajectory, reference evidence, and temporal identity assessment so that a future result can distinguish preservation of a person from faithful execution of that person's requested performance.

A recognizable person can still perform the wrong expression

Identity preservation is often assessed through similarity to a reference, but expressions legitimately change the visible face. A model that remains close to a neutral portrait may receive a favorable similarity score while failing an instruction to laugh, frown, or speak. Conversely, a large facial change may be appropriate to the requested performance. Evaluation should not force these possibilities into a single continuum whose high end is simply resemblance to a static image.

ARGUS conditions video on organized identity observations (Meng et al., 2026b). This provides a useful setting for asking whether identity persists during an intended expression rather than only before or after it. The present article proposes an expression-aware assessment design. It does not claim that ARGUS preserves or loses any particular expression, and it measures no expression trajectory. Its objective is to define the evidence a future study would need before making that combined performance claim.

Describing the requested performance over time

A proposed task would specify an expression or performance with an interpretable temporal boundary. A transition from neutral to smiling is different from a sustained smile, and both differ from speech-associated facial movement. The annotation protocol should reflect those distinctions rather than assign every frame an unrelated expression category. If the instruction leaves timing open, assessment should allow appropriate variation while retaining a clear criterion for whether the requested event occurs.

Reference sets should record the expressions they already contain. A model supplied with the target expression has a different information budget from one asked to produce an unseen expression from neutral references. Both tasks can be legitimate, but their results answer different questions. Matching the number of images does not match their informational content. The study should retain these differences and avoid describing reference reproduction as evidence of the same generalization capability as an unseen performance.

Identity assessment should occur during the expression as well as during easier intervals. If an automated score becomes unreliable under mouth movement or partial visibility, that uncertainty should remain visible. Independent human judgments could assess recognizability with an adequate viewing protocol. They should separately assess expression execution rather than choose a single overall favorite. The resulting outcome table would preserve both dimensions and the cases where the two kinds of success disagree.

Interpreting a joint preservation result

A proposed comparison could use common subjects, instructions, reference evidence, and generated sample counts. Repeated samples should remain linked to their underlying task. A method's best selected expression should not be compared with another method's unselected output unless selection is explicitly part of both workflows. Reporting all attempted generations helps distinguish performance quality from the ability to search for one favorable clip among many attempts.

Temporal aggregation should also be specified in advance. A recognizable beginning and ending cannot establish that the subject remains stable throughout a difficult transition. A worst-frame identity score, however, may respond to one unassessable motion blur. An event-based procedure can record the duration and evidence for suspected substitution while allowing uncertainty. The procedure needs validation against the intended consequence rather than automatic adoption because it yields a convenient single number for each clip.

Finally, the reference population and performance range define the claim's scope. A study of a few common expressions cannot establish preservation under every vocalization, emotional display, or exaggerated movement. Additional conditions can be investigated in later versions using the same explicit records. The proposed design values a bounded positive result: showing that recognizability and a defined performance coexist under documented conditions is more informative than treating a generic similarity average as evidence for all forms of expressive motion.

Measuring identity during an expressive transition

For the proposed comparison, the key observation is expression event. A subject performs a brief exaggerated frown before returning to the neutral look in the reference. This hypothetical case poses a question about recognizability during the changed expression rather than only at the endpoints. Record static neutral look and correct expression at the requested event separately. A stable average identity score can be achieved by skipping the expression entirely. Require event completion first and inspect adjacent windows using both independent observers and visual evidence.

Specify the moment the expression should become visible and give reviewers a short window around it. An acceptable motion may include brief deformation, so define which facial configuration should change and which cues should persist. Compare with a neutral control performed under similar framing. If recognition drops in both conditions, illumination or image quality may be a plausible common factor. If it drops only in the changed interval, inspect the realized expression and the subject evidence together before attributing the effect.

A result about recognizability during the changed expression rather than only at the endpoints needs a denominator that includes all eligible attempts, including ones without a usable score. For the expression event, keep undetectable faces in the log and state which cases remained eligible for recognition.

Store conditioning views, generated frames and visibility annotations separately for expression event. When the same design produces several observations of expression event, start with a group-level contrast and inspect the spread among groups. When examining correct expression at the requested event, inspect pairs of frames with similar pose before drawing conclusions about a persistent identity error.

After evaluating correct expression at the requested event, list the cases whose result changes when the measurement definition is adjusted within a declared range. The evaluation of static neutral look can distinguish an incorrect subject from a permitted change in expression, pose or scene. Set aside examples

of expression event for independent assessment, rather than testing transfer by revisiting the cases that motivated the protocol.

Systems and comparators

Image-conditioned video provides a direct persistence test. Keep the supplied view and later output segment as separate observations. A methodological precedent adds image prompting to a text-conditioned diffusion model (Ye et al., 2023). In an evaluation, a relevant contribution investigates image-to-video generation and video training data curation (Blattmann et al., 2023). Varying the image-conditioning strength gives a practical control for the trade between copying a reference and following the requested transformation.

One approach describes reusable software interfaces for fitting and evaluating estimators (Pedregosa et al., 2011). One study uses a diffusion transformer for text-conditioned video generation (Yang et al., 2024). In an evaluation, a declared software pipeline can preserve the order of transformations and enable held-out validation without refitting preprocessing on the evaluation observations. Use a text-conditioned video baseline to assess generic motion with the same clip duration and language instructions.

An appearance-aware baseline can be compared with motion-only association under the same detections to locate the source of identity changes. An existing method adapts image synthesis to a specific subject using example images (Ruiz et al., 2022). In practical terms, keep reference and evaluation views separate so that recognition after adaptation can be distinguished from successful reproduction of a training portrait. One study uses appearance information in online association (Wojke et al., 2017).

Separate facial structure and texture checks can identify whether a convincing local detail conceals changes in global appearance over time. A separable motion module offers a useful distinction between the visual quality of a still frame and consistency under movement. Earlier research separates global facial structure and local detail in identity-conditioned video (Yuan et al., 2024). During validation, earlier research adds a motion module to personalized image-generation systems (Guo et al., 2023).

Samples and evidence

Documenting construction and intended use clarifies which exclusions occurred during collection, which arose from a model failure and which were imposed during analysis. A relevant contribution defines an unconstrained face verification collection and pair protocol (Huang et al., 2008). In an evaluation, pair definitions determine the verification denominator, so any comparison should state how multiple views of the same person were sampled. A relevant contribution asks dataset creators to record collection, intended use and limitations (Gebru et al., 2021).

Measurement and error

Verification should include genuine and impostor pairs with similar acquisition conditions to avoid turning image quality into a surrogate for identity. Object annotations support localized checks, with the class mapping and annotation policy recorded before evaluating cases absent from the detector vocabulary. Related work provides object annotations and standardized visual recognition tasks (Lin et al., 2014). During validation, earlier research learns an embedding space for face verification and recognition (Schroff et al., 2015).

Related work learns aligned visual and language representations through contrastive training (Radford et al., 2021). A methodological precedent uses an additive angular margin to train discriminative face embeddings (Deng et al., 2018). In an evaluation, a whole-image embedding reflects scene description

and background alongside subject appearance. A local check can distinguish these contributions. Angular discrimination makes a reasonable verification baseline, provided the threshold is evaluated again on profiles, occlusions and synthetic views outside ordinary portrait capture.

Low prevalence can make a seemingly small false-positive rate operationally costly, so report precision at plausible recall levels and retain class counts. Using a representation learned without explicit class labels gives a useful comparison when a prompt-derived score may reward the wrong visual feature. Earlier research contrasts precision-recall and receiver operating characteristic summaries under class imbalance (Saito and Rehmsmeier, 2015). During validation, another study learns visual representations through self-supervision (Caron et al., 2021).

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

Another study constructs a full-reference metric with structure and texture components (Ding et al., 2020). A methodological precedent uses deep features to measure pairwise perceptual resemblance (Zhang et al., 2018). In an evaluation, learned perceptual resemblance responds to some structural differences that pixel errors miss, while remaining dependent on the crop and the chosen comparison target. Separate structure and texture responses are informative when a change in surface detail is permissible but a change in overall shape is consequential.

A paired difference plot can reveal systematic offsets and score-dependent disagreement even when the correlation between two procedures is high. In an evaluation, one approach organizes video-generation evaluation into distinct measurement dimensions (Huang et al., 2023). One study examines discrepancies between paired measurement procedures (Bland and Altman, 1986). Its separate video dimensions provide a reason to publish a failure profile instead of combining every observation into a single quality index.

Linking evidence to the evaluation

A one-reference adapter supplies a baseline for evaluating whether added views help temporal persistence when motion and instructions are held constant. An existing method provides face observations with substantial variation in pose and age (Cao et al., 2017). In practical terms, one published study personalizes video through a face adapter using one reference (He et al., 2024). Pose and age diversity help construct cross-condition verification pairs, but transfer tests still need outputs from the intended workflow. A methodological precedent builds a stacked identity representation from input portraits (Li et al., 2023a). Related work conditions identity-preserving image synthesis on a single reference (Wang et al., 2024).

One study uses auxiliary views to condition identity-preserving video (Wu et al., 2026). Earlier research uses first-frame information to guide spatial and temporal consistency (Ren et al., 2024). Related work tests face recognition with a large distractor collection (Kemelmacher-Shlizerman et al., 2015). In an evaluation, first-frame conditioning gives a clear initial anchor; later tests can distinguish ordinary appearance change from drift after a break in visibility. A multiple-view conditioning baseline is most informative when the views add distinct pose evidence rather than repeated copies of a similar crop. The size and makeup of the distractor gallery alter the false-match burden, so gallery construction belongs in the main experimental description.

Conclusion

Preserving an expressive subject requires observations of both the person and the intended performance over time. This article proposes separate endpoints, documented expression references, and temporal assessment during difficult transitions. It reports no new expression or generation results. A future study using this design could show where identity continuity and expression execution coexist, where they trade off, and where the available measurement evidence remains too weak to decide.

References

Yuan, Shenghai, et al. "Identity-Preserving Text-to-Video Generation by Frequency Decomposition." *arXiv preprint arXiv:2411.17440*, 2024. https://arxiv.org/abs/2411.17440.

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.

Guo, Yuwei, et al. "AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning." *arXiv preprint arXiv:2307.04725*, 2023. https://arxiv.org/abs/2307.04725.

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.

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

Ren, Weiming, et al. "ConsistI2V: Enhancing Visual Consistency for Image-to-Video Generation." *arXiv preprint arXiv:2402.04324*, 2024. https://arxiv.org/abs/2402.04324.

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.

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.

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.

Wu, Mingyang, et al. "ConsID-Gen: View-Consistent and Identity-Preserving Image-to-Video Generation." *arXiv preprint arXiv:2602.10113*, 2026. https://arxiv.org/abs/2602.10113.

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.

Radford, Alec, et al.. "Learning Transferable Visual Models From Natural Language Supervision" *arXiv preprint arXiv:2103.00020*, 2021. https://arxiv.org/abs/2103.00020.

Caron, Mathilde, et al.. "Emerging Properties in Self-Supervised Vision Transformers" *arXiv preprint arXiv:2104.14294*, 2021. https://arxiv.org/abs/2104.14294.

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

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.

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.

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.

Yang, Zhuoyi, et al. "CogVideoX: Text-to-Video Diffusion Models with An Expert Transformer." *arXiv preprint arXiv:2408.06072*, 2024. https://arxiv.org/abs/2408.06072.

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.

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.

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.

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.

Ding, Keyan, et al. "Image Quality Assessment: Unifying Structure and Texture Similarity." *arXiv preprint arXiv:2004.07728*, 2020. https://arxiv.org/abs/2004.07728.

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

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.

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

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.

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.

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.