Separating Identity from Appearance in Subject-Preserving Video

Elijah Barker (CY), Paul Wallace (SK), Austin Dawson (LB)

Abstract

The appearance of a person can change while the intended subject remains the same. This conceptual analysis separates stable identity requirements from requested changes in clothing, expression, lighting, and scene. It proposes paired reference and instruction designs that reveal whether a generator preserves the subject or merely copies familiar visual cues. ARGUS provides the motivating multi-view conditioning approach. No generated clips or recognition results are produced here. The contribution is an evaluation contract that makes permitted appearance variation explicit before a similarity score is interpreted as evidence of identity preservation.

Similarity can reward the wrong invariance

A generated subject may resemble a reference because it copies a distinctive jacket, background, or camera treatment. Those cues can be useful to an image-matching system even when the intended person has changed. Conversely, a requested change of clothing or expression can reduce generic visual similarity while preserving the subject. An identity evaluation should therefore specify what must remain invariant and what the user is allowed to change. Otherwise, the score may reward refusal to perform the requested transformation.

ARGUS organizes multiple identity observations into a structured conditioning memory (Meng et al., 2026b). That design makes it especially useful to ask which visual information the reference collection represents and which aspects an output is expected to preserve. The present article does not infer that its memory necessarily copies or discards particular cues. It develops a conceptual test that could distinguish those outcomes. The focus is the relationship between a preservation claim and a declared appearance transformation, not an unobserved mechanism inside the model.

Defining stable and variable requirements

A proposed study would identify a subject using an appropriate reference set and then specify a transformation whose success can be assessed independently of subject identity. The transformation might alter the scene or an allowed appearance attribute. The identity endpoint would ask whether the output remains recognizable as the intended subject under that change. The transformation endpoint would ask whether the requested change occurred. A favorable result on one endpoint should not be used to assume the other.

Reference access should be documented carefully. A model supplied with a view that already contains the requested appearance may face a different task from one asked to combine identity with a genuinely new condition. This does not make the first workflow illegitimate, but it changes the claim. The protocol should record which reference observations contain which cues and match that access across compared systems where feasible. An unexplained difference in reference information can otherwise be mistaken for an architectural advantage.

An independent assessment should also retain cases where identity cannot be judged confidently. A profile, an occluded face, or a highly stylized image may exceed the evaluator's supported conditions. Dropping those cases can inflate apparent preservation, while forcing a decision can confuse measurement error with generator error. The report should provide coverage and preserve ambiguous observations for a

separate analysis. The evaluation contract should say whether the intended identity claim concerns human recognizability, automated verification, or another explicitly measured outcome.

Testing copying and preservation together

A useful diagnostic design would vary reference context while keeping the subject information as comparable as possible. It could ask whether a background or accessory travels into the output when the requested scene changes. It would also include conditions in which copying that cue would be correct, so the evaluator does not treat every reference resemblance as a defect. The scientific question is whether the generator uses the right invariances for the stated task. The experiment needs controlled observations to establish that distinction.

The same logic applies to output diversity. Several valid appearances can belong to the same subject. A low-variation generator may score well by staying near the reference, but provide little of the requested flexibility. A highly varied generator may drift away from the intended person. A meaningful profile would show the identity-transformation relationship across repeated outputs, rather than select one attractive example or collapse both properties into an unexplained total. No such output distribution is measured in this conceptual article.

Finally, the intended use constrains which changes matter. A cinematic workflow may value expressive and stylistic flexibility; another task may require a narrow, consistent reference appearance. The evaluation should describe that use and avoid making a universal identity claim from one preferred setting. A finite reference set cannot establish every possible appearance of a person. The goal is a bounded preservation result supported by the actual views and transformations that the study observes.

A same-person wardrobe and scene crossover

For the proposed comparison, the key observation is reference-to-clip comparison. The source portrait shows a red coat in a station while the generated scene calls for a white uniform in a kitchen. This hypothetical case poses a question about recognizability conditional on completed appearance change. Record identity preserved with requested change and reference appearance copied with scene details separately. A background-sensitive embedding could reward a visually similar station clip even when its person is different. Have reviewers judge person and requested appearance separately, using a same-outfit distractor to check their criteria.

Place the red-coat distractor next to the same person in the new wardrobe. The difficult decision is whether the evaluator follows the person when the more obvious visual match belongs to someone else. Obtain a separate judgment of the new scene, because a portrait copied into the original station never completed the transformation. If the evaluations disagree, inspect the cropped face and the full clip in separate conditions. That contrast can help identify which visible cues account for the discrepancy.

Before estimating recognizability conditional on completed appearance change, fix the conditions under which a case counts as eligible and decide how exclusions will appear in the summary. For the reference-to-clip comparison, keep undetectable faces in the log and state which cases remained eligible for recognition.

Store conditioning views, generated frames and visibility annotations separately for reference-to-clip comparison. When the same design produces several observations of reference-to-clip comparison, start with a group-level contrast and inspect the spread among groups. When examining reference appearance copied with scene details, inspect pairs of frames with similar pose before drawing conclusions about a persistent identity error.

Use the comparison with identity preserved with requested change to identify which parts of the finding survive a reasonable change in the decision threshold. The evaluation of identity preserved with requested change can distinguish an incorrect subject from a permitted change in expression, pose or scene. Set aside examples of reference-to-clip comparison for independent assessment, rather than testing transfer by revisiting the cases that motivated the protocol.

Systems and comparators

An appearance-aware baseline can be compared with motion-only association under the same detections to locate the source of identity changes. A methodological precedent learns concept embeddings from a few images (Gal et al., 2022). For a prospective comparison, related work uses appearance information in online association (Wojke et al., 2017). A learned concept embedding offers a compact control for comparing changes in prompt adherence under the same set of reference images.

An adapter can hold the type of external condition constant when comparing how faithfully different systems honor an edge, depth or pose specification. Multiple concept inputs expose cases in which both people appear correctly but their assigned roles are reversed. During validation, related work introduces adapters to convey spatial or visual conditions to image diffusion (Mou et al., 2023). A methodological precedent adapts diffusion conditioning for multiple concepts (Kumari et al., 2022).

One approach uses tuning-free identity customization with attention to preserving image editability (Guo et al., 2024). For a prospective comparison, adaptation comparisons can report trainable parameters together with sample count and compute, keeping the data budget distinct from the architectural change. A paired edit test can separate matching the reference subject from following a permitted change to composition, clothing or surrounding scene. One study adapts pretrained models through low-rank parameter updates (Hu et al., 2022).

Another study adds spatial controls to pretrained text-to-image diffusion (Zhang et al., 2023). An existing method adapts image synthesis to a specific subject using example images (Ruiz et al., 2022). Fixed spatial controls offer a way to vary prompts while holding a geometric scaffold stable, or to vary the scaffold under common language. In a controlled test, keep reference and evaluation views separate so that recognition after adaptation can be distinguished from successful reproduction of a training portrait.

Varying the image-conditioning strength gives a practical control for the trade between copying a reference and following the requested transformation. Image-conditioned video provides a direct persistence test. Keep the supplied view and later output segment as separate observations. During validation, one study investigates image-to-video generation and video training data curation (Blattmann et al., 2023). An existing method adds image prompting to a text-conditioned diffusion model (Ye et al., 2023).

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

Measurement and error

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

A cosine-margin baseline helps distinguish rejection due to the representation from rejection caused by the threshold. One published study investigates scaling in contrastive language-image learning (Cherti et al., 2022). For a prospective comparison, representation changes can alter relative prompt scores, so the encoder checkpoint and text preprocessing belong in the comparison record. An existing method uses a cosine margin for discriminative face recognition (Wang et al., 2018).

Verification should include genuine and impostor pairs with similar acquisition conditions to avoid turning image quality into a surrogate for identity. One approach uses multiscale image representations for quality prediction (Ke et al., 2021). A multiscale score can help distinguish sensitivity to resampling or crop size from a stable image-level judgment. One approach learns an embedding space for face verification and recognition (Schroff et al., 2015).

One study learns general-purpose visual features from curated image data (Oquab et al., 2023). A quality-sensitive representation can be used to define inspectable strata before comparing acceptance rates across views with very different visibility. In practical terms, a general visual representation can provide an independent similarity score whose errors may differ from those of a face or prompt-specific model. A relevant contribution learns facial representations with a quality-related feature magnitude (Meng et al., 2021).

A relevant contribution constructs a full-reference metric with structure and texture components (Ding et al., 2020). A methodological precedent uses an additive angular margin to train discriminative face embeddings (Deng et al., 2018). Separate structure and texture responses are informative when a change in surface detail is permissible but a change in overall shape is consequential. For a prospective comparison, angular discrimination makes a reasonable verification baseline, provided the threshold is evaluated again on profiles, occlusions and synthetic views outside ordinary portrait capture.

Samples and evidence

Related work defines an unconstrained face verification collection and pair protocol (Huang et al., 2008). An existing method tests face recognition with a large distractor collection (Kemelmacher-Shlizerman et al., 2015). Pair definitions determine the verification denominator, so any comparison should state how multiple views of the same person were sampled. In an evaluation, the size and makeup of the distractor gallery alter the false-match burden, so gallery construction belongs in the main experimental description.

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

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. A multiple-view conditioning baseline is most informative when the views add distinct pose evidence rather than repeated copies of a similar crop. In practical terms, related work uses auxiliary views to condition identity-preserving video (Wu et al., 2026). A methodological precedent conditions identity-preserving image synthesis on a single reference (Wang et al., 2024). One approach personalizes video through a face adapter using one reference (He et al., 2024).

Conclusion

Identity preservation should not be equated automatically with global appearance similarity. A useful evaluation states which subject properties must persist, which changes are requested, and which reference cues are available. It then observes preservation and transformation as separate outcomes, with measurement coverage retained. The proposed contract would make a future ARGUS-related comparison more interpretable without presuming its result. A generator demonstrates the relevant capability when it preserves the intended subject through the changes the task actually requires.

References

Guo, Zinan, et al. "PuLID: Pure and Lightning ID Customization via Contrastive Alignment." *arXiv preprint arXiv:2404.16022*, 2024. https://arxiv.org/abs/2404.16022.

Zhang, Lvmin, et al. "Adding Conditional Control to Text-to-Image Diffusion Models." *arXiv preprint arXiv:2302.05543*, 2023. https://arxiv.org/abs/2302.05543.

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

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.

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.

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

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, 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.

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.

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

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.

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

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

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.

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.

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.

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.

Ke, Junjie, et al.. "MUSIQ: Multi-scale Image Quality Transformer" *arXiv preprint arXiv:2108.05997*, 2021. https://arxiv.org/abs/2108.05997.

Meng, Qiang, et al. "MagFace: A Universal Representation for Face Recognition and Quality Assessment." *arXiv preprint arXiv:2103.06627*, 2021. https://arxiv.org/abs/2103.06627.

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.

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.

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.

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

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.

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.

Kumari, Nupur, et al. "Multi-Concept Customization of Text-to-Image Diffusion." *arXiv preprint arXiv:2212.04488*, 2022. https://arxiv.org/abs/2212.04488.

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

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.

Mou, Chong, et al. "T2I-Adapter: Learning Adapters to Dig out More Controllable Ability for Text-to-Image Diffusion Models." *arXiv preprint arXiv:2302.08453*, 2023. https://arxiv.org/abs/2302.08453.

Hu, Edward J., et al. “LoRA: Low-Rank Adaptation of Large Language Models.” *International Conference on Learning Representations*, 2022. https://openreview.net/forum?id=nZeVKeeFYf9.