Texture Fidelity and Recognizability under Reference Transfer

Jason Reed (PA), Kenneth Jenkins (BN), Arthur Jenkins (JM)

Abstract

Fine visual detail can improve reference resemblance without settling whether a generated person remains recognizable. This conceptual study separates texture fidelity, subject identity, and the quality of the requested transformation. It proposes matched-resolution reference comparisons and independent assessment of their output consequences. ARGUS motivates the structured-reference setting, but no texture manipulation or generation experiment is performed. The argument is that an identity-preservation claim should not inherit its meaning from whichever visual frequency or image representation a convenient similarity metric happens to emphasize.

Detail can matter in different ways

A reference contains both broad structure and fine appearance. Some detail may contribute to recognizing a subject; other detail may reflect lighting, compression, camera processing, or a transient surface condition. A generator can reproduce these properties with different degrees of success. An evaluation should state which visual resemblance is relevant to the identity task rather than treat every copied texture as equal evidence of preservation.

ARGUS uses structured identity references to condition video (Meng et al., 2026b). This motivates asking how the quality and resolution of reference evidence influence the resulting subject, without assuming that its mechanism relies on a particular frequency band. The present article develops a conceptual measurement design. It does not modify reference images, inspect feature maps, or calculate new identity scores. Its purpose is to separate a claim about fine detail from a claim about recognizable subject continuity under a requested change.

Matching resolution and documenting transformations

A proposed experiment could vary reference resolution or a declared detail transformation while preserving the underlying source observation. The processing pipeline would need to be recorded. Resampling can alter boundaries, local contrast, and compression artifacts, so it is not necessarily a pure intervention on identity information. Independent checks should establish which evidence remains available. A lower preservation score after processing would not by itself prove that the original system used a specific fine-detail cue.

The generator and evaluator should receive clearly defined versions of the reference. If both change simultaneously, a score difference can reflect either the produced output or the measurement instrument. A useful design could hold generated clips fixed while changing evaluation references, then hold the evaluator fixed while changing generation references. These are separate comparisons with different interpretations. A complete workflow result can still be reported, but it should not be presented as an isolated mechanism finding without the relevant controls.

Output endpoints should include recognizability and the intended transformation. A model might preserve high-frequency detail by keeping pose, expression, or lighting close to the reference. If the task requires a different condition, that strategy may sacrifice instruction execution. Conversely, a valid change can alter texture while retaining the subject. A study should retain those disagreements and avoid selecting a single similarity score whose preferred invariance predetermines the winner.

Interpreting detail across a video sequence

Video adds a temporal dimension to texture fidelity. A detail can appear sharp in individual frames while varying inconsistently over time. A temporally stable output can instead smooth away details that an observer finds important. The assessment should preserve both local and sequence-level observations, with a declared consequence rule. An average over frames cannot establish the persistence of every feature, while a worst-frame result can be driven by an unreliable measurement during motion or occlusion.

Independent human judgments may help connect a numerical resemblance to recognizable identity, but their presentation matters. Observers need appropriate references and task instructions, and uncertainty should be retained where the visible evidence is inadequate. A person asked only to compare image detail may answer a different question from one asked whether the subject remains recognizable through an unfamiliar expression. Those judgments should not be merged automatically under a general label of fidelity.

Finally, the evaluation should state its source population. A result using high-quality reference photographs may not describe low-resolution or compressed inputs from another workflow. Reweighting can change emphasis among observed cases but cannot supply missing evidence. The proposed study would report the reference conditions actually supported, along with failures to generate or assess. A bounded result can be useful without implying that texture preservation has been established across every acquisition and output condition.

Local detail after a matched change of view

For the proposed comparison, the key observation is matched output and reference region. A close view contains a crisp facial texture but a wider view changes the proportions that make the subject recognizable. This hypothetical case poses a question about independently judged recognizability under equal task completion. Record sharp local features and stable global appearance under transformation separately. Texture copied from the input can inflate a perceptual similarity score even when the broader appearance is inconsistent. Review region-level detail and person-level recognition separately before interpreting their agreement.

Collect separate judgments at two magnifications using the same output frame and view. Fine texture may be recoverable in a crop even when the full view makes an altered facial shape visible. A reference patch copied from the original can be tested as a deliberately local control. Compare whether detail scores track the patch and whether observers still recognize the person in context. Keep the requested transformation as its own outcome so that an exact reference crop cannot score as a successful new view on its own.

Before estimating independently judged recognizability under equal task completion, fix the conditions under which a case counts as eligible and decide how exclusions will appear in the summary. For the matched output and reference region, keep undetectable faces in the log and state which cases remained eligible for recognition.

Store conditioning views, generated frames and visibility annotations separately for matched output and reference region. If there are many observations of matched output and reference region, decide which ones share an underlying history and keep that grouping in any uncertainty analysis. When examining stable global appearance under transformation, inspect pairs of frames with similar pose before drawing conclusions about a persistent identity error.

A sensitivity display can show where the conclusion about independently judged recognizability under equal task completion changes as assumptions about the eligible observations change. The evaluation of sharp local features can distinguish an incorrect subject from a permitted change in expression, pose or scene. A transfer study of independently judged recognizability under equal task completion would need fresh cases chosen independently of the observed failures.

Systems and comparators

A one-reference adapter supplies a baseline for evaluating whether added views help temporal persistence when motion and instructions are held constant. Separate facial structure and texture checks can identify whether a convincing local detail conceals changes in global appearance over time. Earlier research personalizes video through a face adapter using one reference (He et al., 2024). In practical terms, one published study separates global facial structure and local detail in identity-conditioned video (Yuan et al., 2024).

An existing method investigates image-to-video generation and video training data curation (Blattmann et al., 2023). Another study describes reusable software interfaces for fitting and evaluating estimators (Pedregosa et al., 2011). For implementation, a declared software pipeline can preserve the order of transformations and enable held-out validation without refitting preprocessing on the evaluation observations. Image-conditioned video provides a direct persistence test. Keep the supplied view and later output segment as separate observations.

A separable motion module offers a useful distinction between the visual quality of a still frame and consistency under movement. In an evaluation, a paired edit test can separate matching the reference subject from following a permitted change to composition, clothing or surrounding scene. Related work adds a motion module to personalized image-generation systems (Guo et al., 2023). Earlier research uses tuning-free identity customization with attention to preserving image editability (Guo et al., 2024).

Measurement and error

Another study uses deep features to measure pairwise perceptual resemblance (Zhang et al., 2018). In an evaluation, feature magnitude and recognition confidence can help identify low-quality crops, but agreement with an independent quality annotation should be checked first. Learned perceptual resemblance responds to some structural differences that pixel errors miss, while remaining dependent on the crop and the chosen comparison target. Another study uses feature norms to adapt training margins to face image quality (Kim et al., 2022).

One study distinguishes forms of intraclass correlation according to the rating design (Koo and Li, 2016). For a prospective comparison, one published study learns facial representations with a quality-related feature magnitude (Meng et al., 2021). A quality-sensitive representation can be used to define inspectable strata before comparing acceptance rates across views with very different visibility. Report which raters, measurements and agreement definition enter the coefficient, since different designs give superficially similar statistics different meanings.

One study examines discrepancies between paired measurement procedures (Bland and Altman, 1986). A multiscale score can help distinguish sensitivity to resampling or crop size from a stable image-level judgment. A paired difference plot can reveal systematic offsets and score-dependent disagreement even when the correlation between two procedures is high. During validation, one study uses multiscale image representations for quality prediction (Ke et al., 2021).

Scaling choices determine how units of different magnitude contribute to an aggregate error; a transparent report retains the original-scale results where interpretable. One study assesses probability calibration and studies temperature scaling (Guo et al., 2017). In an evaluation, one approach compares choices of forecast error scale across series (Hyndman and Koehler, 2006). After fitting a decision threshold, reliability curves on an independent split can show whether reported confidence still matches observed success frequency.

Local structural agreement can be inspected alongside task metrics when pixel changes are expected but geometry or regional coherence should remain stable. Its separate video dimensions provide a reason to publish a failure profile instead of combining every observation into a single quality index. In practical terms, an existing method compares the local structure of a pair of images (Wang et al., 2004). Related work organizes video-generation evaluation into distinct measurement dimensions (Huang et al., 2023).

Landmark localization provides a way to identify crop failures before measuring resemblance and to retain a denominator for faces that were never detected. Confidence calibrated on the original collection may change under a new capture or generation regime, so calibration belongs in transfer testing. Earlier research jointly localizes faces and facial landmarks (Deng et al., 2019). In practical terms, an existing method examines uncertainty estimates under dataset shift (Ovadia et al., 2019).

Separate structure and texture responses are informative when a change in surface detail is permissible but a change in overall shape is consequential. Verification should include genuine and impostor pairs with similar acquisition conditions to avoid turning image quality into a surrogate for identity. One published study learns an embedding space for face verification and recognition (Schroff et al., 2015). For a prospective comparison, a methodological precedent constructs a full-reference metric with structure and texture components (Ding et al., 2020).

Analysis and interpretation

A relevant contribution investigates reliable comparisons of reinforcement learning under limited runs (Agarwal et al., 2021). In an evaluation, related work conditions identity-preserving image synthesis on a single reference (Wang et al., 2024). Uncertainty intervals and performance profiles help show whether an apparent advantage persists across titles with very different return distributions. Single-reference conditioning provides a budget-matched comparison for identifying which errors disappear when extra views are made available to the generator.

Samples and evidence

Pose and age diversity help construct cross-condition verification pairs, but transfer tests still need outputs from the intended workflow. A methodological precedent defines an unconstrained face verification collection and pair protocol (Huang et al., 2008). A relevant contribution provides face observations with substantial variation in pose and age (Cao et al., 2017). In practical terms, pair definitions determine the verification denominator, so any comparison should state how multiple views of the same person were sampled.

Linking evidence to the evaluation

Related work builds a stacked identity representation from input portraits (Li et al., 2023a). Related work uses auxiliary views to condition identity-preserving video (Wu et al., 2026). One approach adapts image synthesis to a specific subject using example images (Ruiz et al., 2022). A multiple-view conditioning baseline is most informative when the views add distinct pose evidence rather than repeated copies of a similar crop. Keep reference and evaluation views separate so that recognition after adaptation can be distinguished from successful reproduction of a training portrait. During validation, a stacked portrait input permits a test of view coverage, provided the evaluator retains an independent portrait and records near-duplicate input views.

The size and makeup of the distractor gallery alter the false-match burden, so gallery construction belongs in the main experimental description. For implementation, a reproducible result requires identifiers for both the released inputs and the exact intermediate tables from which the paper derives its figures. Another study specifies principles for discoverable and reusable research artifacts (Wilkinson et al., 2016). A relevant contribution tests face recognition with a large distractor collection (Kemelmacher-Shlizerman et al., 2015).

Conclusion

Texture fidelity should be connected explicitly to the identity and transformation requirements of the task. The proposed design separates reference processing, generation, evaluation, and temporal appearance, while retaining ambiguous judgments and coverage. It reports no new performance observation. Its contribution is to make fine-detail evidence interpretable without allowing generic reference resemblance to substitute for the particular subject-preservation capability a video study intends to establish.

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.

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

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.

Kim, Minchul, et al. "AdaFace: Quality Adaptive Margin for Face Recognition." *arXiv preprint arXiv:2204.00964*, 2022. https://arxiv.org/abs/2204.00964.

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.

Koo, Terry K., and Mae Y. Li. "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research." *Journal of Chiropractic Medicine*, vol. 15, no. 2, 2016, pp. 155-163. https://doi.org/10.1016/j.jcm.2016.02.012.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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

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

Hyndman, Rob J. and Anne B. Koehler. “Another look at measures of forecast accuracy.” *International Journal of Forecasting*, vol. 22, no. 4, 2006, pp. 679-688. https://doi.org/10.1016/j.ijforecast.2006.03.001.

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.

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.

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.

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

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.

Agarwal, Rishabh, et al.. "Deep Reinforcement Learning at the Edge of the Statistical Precipice" *arXiv preprint arXiv:2108.13264*, 2021. https://arxiv.org/abs/2108.13264.

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.

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

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.

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.

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.

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.