PULSE AI WORKING PAPER · VERSION 1.0 · JULY 2026

Identity Persistence in AI-Generated Video

A Sequence Evaluation Framework for Founder-Led Outbound

Ananya Das

Pulse AI, India · pulseai.pro

ORCID: 0009-0007-2490-8231

ABSTRACT

AI video tools are built and benchmarked to produce one good clip. Founder-led outbound does not sell in one clip — it sells across a sequence, and the sender is the trust signal carrying it. When the face or visual style shifts between the first outreach and the follow-up, the recipient stops believing the message came from a person and stops replying; the commercial cost of that shift is unmeasured because no existing benchmark looks across outputs. This paper defines that failure mode as *identity drift* and specifies an evaluation framework that treats an outbound sequence, not a single video, as the unit of analysis. Three metrics are defined: face persistence, style persistence, and a composite Sequence Persistence Score (SPS). A sequence passes only if the sender remains recognisably the same person across every moment in it. The generation pipeline is proprietary; the evaluation protocol is published in full so that any comparable pipeline can be scored against the same thresholds.

Keywords: identity drift; AI-generated video; face embeddings; sequence persistence; evaluation framework; founder-led outbound

AUTHOR'S NOTE

I am building the system this framework evaluates, and I publish the evaluation because the product is the claim and the measurement is the proof. My conviction is that the next phase of outbound is not volume but persistent generative identity — a sender who stays the same person across every message, at scale. That claim is worth making only if it can be checked, which is what this protocol is for. It is published in a form that lets a reader test my system, or a competing one, against the same standard.

What This Argument Rests On

Three claims, stated plainly so a reader can reject them early:

- **B2B sales is a sequence, not a message.** High-consideration deals close across multiple touches. Any evaluation that scores one output in isolation is scoring the wrong unit.
- **Identity drift breaks trust before it breaks quality.** A recipient does not consciously assess video fidelity; they assess whether the sender is a real, consistent person. Drift fails that assessment while every single-clip quality metric still passes.
- **Persistence is measurable, and measurement is the differentiator.** Single-output quality is the wrong acceptance criterion for relationship-led sales. The platforms that win this category will be the ones that can demonstrate consistency across a sequence rather than assert it — which requires a published, checkable standard.

If those three hold, the rest of this paper is the instrument. If any of them fails, the framework is unnecessary.

1. Problem Definition

We call the failure mode **identity drift**: the gradual loss of face, style, and framing consistency across multiple AI-generated videos of the same person.

In founder-led outbound, the sender is the trust signal. When a prospect receives three or four video messages, they must all feel like they came from the same person. If the face changes or the visual style jumps between clips, the recipient stops trusting the message and stops replying.

The published benchmarks for AI video optimise for the wrong unit. They score one clip in isolation. None of them asks whether the sender survives the sequence — which is the only question a prospect on the fourth touch is actually answering. That is the gap this framework closes.

2. Related Work

Face recognition embeddings provide the identity signal used here; we adopt ArcFace [1] via the InsightFace implementation [2]. The generation side draws on image animation [3] and audio-driven lip synthesis [4], both of which are evaluated in the literature on per-output realism rather than on consistency across a set of outputs.

Forensics already treats identity inconsistency as signal — it is how manipulated media gets caught [5]. This framework inverts that use. The same measurement that detectors run after the fact becomes a production gate applied before a sequence ships, and it runs across separate generated artifacts rather than within a single one. No prior work formalises sequence-level identity persistence across independently generated videos of one sender as an evaluation target; that is the gap this paper claims.

3. Methodology

The test setup treats a sequence as the unit of analysis, not a single video.

3.1 Sample selection

- One consented sender identity asset (portrait photo).
- A set of real outbound scripts for a founder-led sales sequence: first outreach, follow-up, post-call reminder.
- One script per sequence moment.

3.2 Pipeline

Every video is generated through the same production pipeline: text-to-speech converts script to audio; an image-to-video model drives body motion from the static portrait; an audio-driven lip-sync stage aligns mouth movement; local enhancement applies motion interpolation and sharpening. Pipeline configuration is held constant across every clip in a sequence.

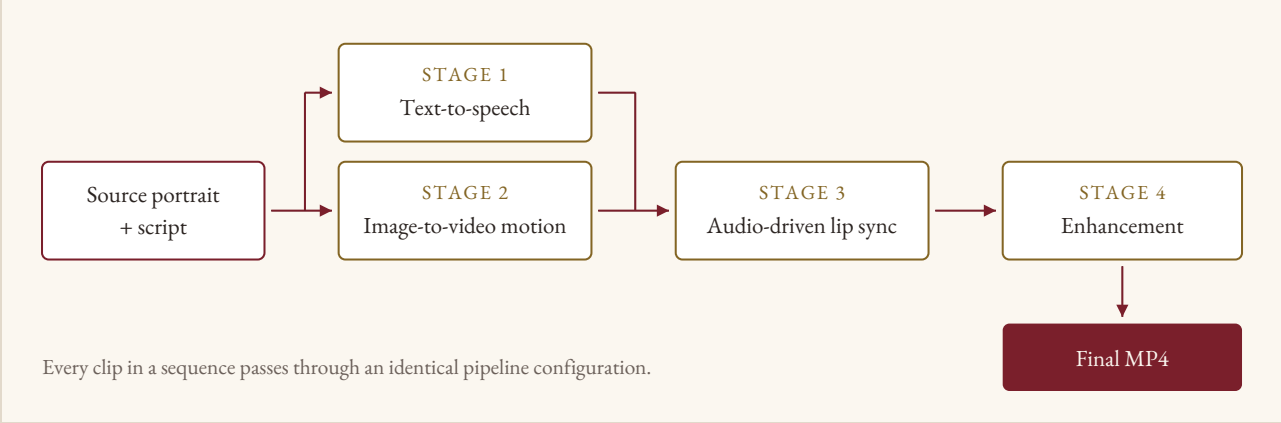

Figure 1. Generation pipeline. In founder-led outbound, visual variation between messages is not a creative choice; it is a trust-destroying defect. Holding the pipeline constant ensures that any deviation is attributable to the model rather than to the prompt.

3.3 Annotation process

Every generated video is compared against two references: the original source portrait (source-to-output identity) and the previous video in the sequence (output-to-output persistence). Annotations use face-embedding cosine similarity and style feature distance. Human spot-checks are specified for cases where automated scoring is ambiguous or a clip sits near a threshold; no systematic human review has been carried out at this sample size.

4. Evaluation Metrics

A sequence passes only if the sender remains recognisably the same across all moments.

METRIC	WHAT IT CHECKS	PASS THRESHOLD
Face persistence	Cosine similarity of face embeddings between the source photo and each output, and between consecutive outputs	> 0.65
Style persistence	Distance in background, lighting, and colour-grade features between outputs *(extractor not finalised — see 4.4)*	provisional
Sequence Persistence Score	Weighted mean of face and style persistence across the full sequence ($w_f = 0.6$,	> 0.70

$w_s = 0.4$; see 4.2)

SPS is the headline number and the gate. A sequence that clears it is one the recipient can read as coming from a single person; a sequence below it is regenerated or manually reviewed before it ever reaches a prospect. The threshold is an engineering commitment, not an observed correlation with recipient behaviour — that correlation is the open question in Section 5.

4.1 Embedding configuration

Face embeddings are extracted with InsightFace `buffalo_l` (ArcFace R100) using its bundled detector and 112×112 alignment. Embeddings are L2-normalised and we report cosine similarity, where higher values indicate greater persistence.

The source asset is a single portrait photograph and embeds directly. Each output is a video and must be reduced to one vector before comparison: we sample five evenly spaced frames across the clip, embed each independently, take the arithmetic mean of the five vectors, and re-normalise. Where a frame yields no detected face it is skipped and logged; where it yields several, the highest-confidence detection is used. A clip with fewer than three usable frames is excluded rather than scored.

Frame averaging is a deliberate simplification and it carries a cost: identity variation *within* a single clip averages toward the mean and can pass a comparison that per-frame scoring would fail. Per-frame variance is logged alongside each clip embedding but is not currently part of the score.

Pair counting is as follows. For a sequence of n outputs, face persistence is computed over n source-to-output pairs and $n - 1$ consecutive output-to-output pairs. Pairs are unordered and each is counted once; non-consecutive pairs are not scored.

4.2 Sequence Persistence Score

SPS is computed as a weighted mean of the two component scores across a sequence:

$$\mathbf{SPS} = \mathbf{w_f} \cdot \bar{\mathbf{F}} + \mathbf{w_s} \cdot \bar{\mathbf{S}} \quad \text{where } w_f = 0.6,\ w_s = 0.4$$

$\bar{F}$ is the mean face-persistence similarity across all source-to-output and consecutive output-to-output pairs in the sequence; $\bar{S}$ is the mean style-persistence score across consecutive outputs. Face persistence carries the higher weight because style variation occurs naturally in authentic founder video — different room, different lighting, different week — while face variation never does; weighting style equally would penalise the system for reproducing what genuine sequences look like. This weighting is a design choice, not an empirical result; see Sections 4.3 and 7.

Worked example. A three-clip sequence yields source-to-output similarities of 0.94, 0.91, and 0.93, and consecutive output-to-output similarities of 0.96 and 0.92. Face persistence is the mean of all five values: $\bar{F} = 0.932$. Every pair clears the 0.65 floor, so the sequence is not rejected on face grounds. With a style score of $\bar{S} = 0.70$, SPS = 0.6 × 0.932 + 0.4 × 0.70 = 0.839, which clears the 0.70 gate and ships. Had the third clip drifted to 0.62, that pair would breach the floor and the sequence would be flagged regardless of its composite score.

4.3 Status of the style-persistence component

Style persistence is specified in intent but not in implementation. The feature extractor, the distance function, and the clustering procedure behind "within same cluster" are not yet fixed, and no threshold has been set. This is a material gap rather than a detail: since SPS weights style at 0.4, two fifths of the headline score currently rests on an unspecified quantity. Readers reproducing this protocol should treat the face-persistence half as fully specified and the style half as a placeholder pending v1.1.

4.4 Status of the 0.65 value

The 0.65 face-persistence value functions as an acceptance floor rather than a discrimination boundary: it flags catastrophic identity failure, not gradual drift. Sequences are expected to score well above it. It has not yet been calibrated against a measured different-identity distribution; that calibration is deferred to v1.1.

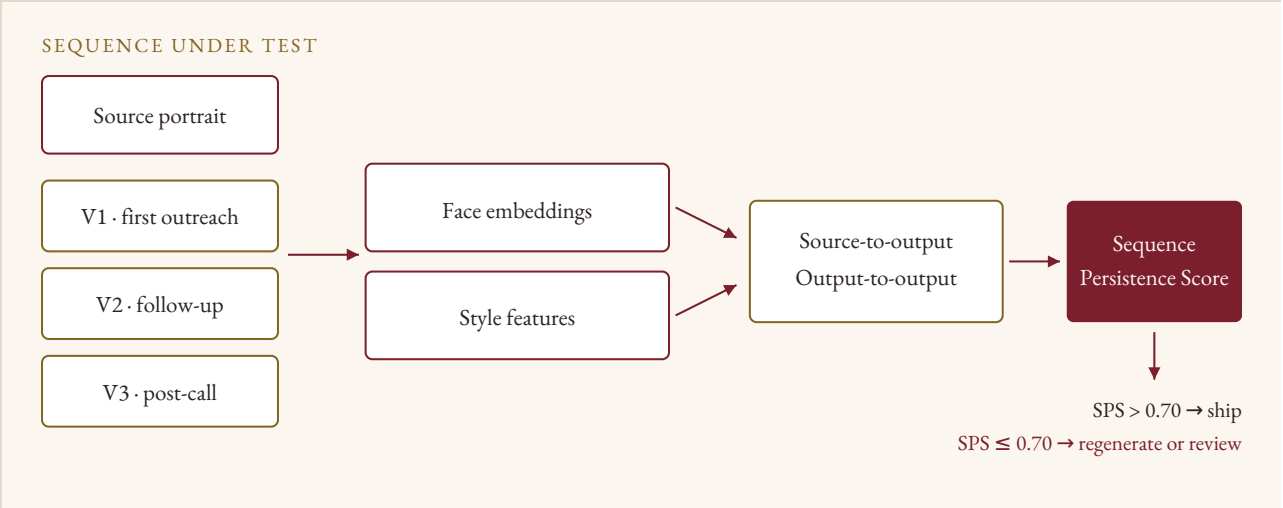

Figure 2. Evaluation flow. Face and style features are extracted per clip, compared against both the source portrait and the preceding clip, then combined into a single pass/fail score.

5. Results Summary

This is a working baseline, not a peer-reviewed study. Results to date come from a single-subject feasibility study: one consented founder identity across three to five sequence moments, run to establish measurement validity rather than to support a generalizable claim. Multi-subject evaluation is deferred to v1.1. Four patterns are observed:

- Per-clip output quality was not a limiting factor in any run; the failures of interest occur between clips, not within them.
- Face persistence across three to five videos is consistently high when the source photo and pipeline are held constant; the observed range is reported in v1.1 once multi-identity calibration is complete.
- Style drift appears qualitatively when the pipeline is varied between runs; this has not yet been quantified, as the style-feature extractor is not finalised.
- By construction, SPS is a stricter gate than either component alone: a sequence can clear the face threshold and still fail the composite. Whether that strictness is correctly calibrated is untested.

The operating hypothesis is that sequences falling below the SPS threshold suffer measurably lower reply and meeting-book rates, because the trust signal the recipient is reading breaks before the argument does. That hypothesis is untested. No outcome data has been collected, and nothing here should be read as evidence that persistence scores predict recipient behaviour.

Testing it requires live sequences and consented outcome data. Founders running high-touch outbound who want their sequences measured against this framework — and their outcome data included in the correlation study — can reach the author via pulseai.pro.

A second open question is cost. The economic burden of the gate is regeneration, not evaluation: embedding extraction across a sequence is computationally negligible, and pipeline-constancy enforcement is a configuration constraint rather than additional compute. The operative figure is the proportion of sequences that fail the threshold and must be regenerated, which sets the cost multiple over single-pass generation. That rate has not been measured and cannot be estimated from the runs completed to date.

6. Reproducibility

The full Pulse AI generation pipeline is proprietary. Reproducibility is therefore provided through a transparent evaluation protocol rather than a public production repository. Any reader can reimplement the Sequence Persistence evaluation from the steps below.

1. Select one consented sender identity asset (portrait photo) and a founder-led outbound sequence: first outreach, follow-up, post-call reminder.
2. Generate the sequence using any image-to-video and lip-sync pipeline that preserves identity, followed by the same light post-processing (motion interpolation and sharpening) on every clip.
3. Extract face embeddings with InsightFace `buffalo_l` (ArcFace R100), L2-normalised, and style features covering background, lighting, and colour grade for each output. Report cosine similarity, where higher values indicate greater persistence.
4. Compute source-to-output and output-to-output cosine similarities.
5. Apply the pass thresholds in Section 4 and compute the Sequence Persistence Score. Any sequence below 0.70 is flagged for regeneration or manual review.

Code and data availability

The standalone evaluation repository is public at github.com/byananya/sequence-persistence-evaluation, released under the MIT licence. It contains the embedding and style-feature extraction code, the SPS computation, unit tests for the metric functions, a Colab-ready reproduction notebook, and diagram sources. It is published so that the framework can be run against any pipeline, including this one: take whichever AI video tool you currently use, run a sequence through the notebook, and read the score. A pipeline that clears the threshold needs nothing from us. Repository contents at the time of writing correspond to commit history on `main`; the archived release accompanying this version is the authoritative snapshot. A consented sample dataset will be added once multi-subject evaluation is carried out. For early access or collaboration, use the contact address listed at pulseai.pro.

7. Limitations

- Sample size is one consented identity; no multi-subject evaluation has been carried out.
- The 0.65 face-persistence value is an uncalibrated acceptance floor, not a measured decision boundary; no different-identity distribution has yet been established against which to place it.
- The face-to-style weighting in the Sequence Persistence Score has not been validated by ablation or human preference data.
- No independent validation of the metric has been performed.
- No claim is made of sales lift, reply-rate improvement, or benchmark superiority over existing systems.
- Style-feature clustering is sensitive to the chosen feature extractor; cross-implementation comparability is untested.

8. Planned Work

The gaps disclosed above define the next version rather than sitting as open caveats. Version 1.1 is scoped to five items, in dependency order:

1. **Calibrate the face-persistence threshold.** Measure a different-identity similarity distribution on this exact embedding configuration, report n, mean, and standard deviation for both the negative and same-source sets, and place the threshold by measured separation rather than by assertion.
2. **Specify the style-persistence component.** Fix the feature extractor, the distance function, and the clustering procedure; set and justify a threshold. Until this lands, two fifths of SPS is unspecified.
3. **Extend to multiple identities.** Move from a single-subject feasibility study to a sample large enough to report a distribution rather than a set of readings.
4. **Measure the regeneration rate.** Establish what proportion of sequences fail the gate, which determines the cost multiple over single-pass generation.
5. **Test the outcome hypothesis.** Correlate SPS against reply and meeting-book rates on live consented sequences, with the explicit possibility that no correlation exists.

Items 1 and 2 are prerequisites for the rest: neither multi-subject evaluation nor outcome correlation is meaningful while the threshold is uncalibrated and the style component undefined.

9. Commercial Implications

Not every outbound programme needs this. If the objective is volume — ten thousand generic videos, no continuity between touches, no expectation that a recipient will receive more than one — existing single-output tools are adequate and this framework adds cost without benefit.

The framework applies where the sender is the asset: founder-led sequences, relationship-led sales, any programme where the same recipient sees the same person more than once and the reply depends on believing it is that person. In that setting, single-output quality is the wrong acceptance criterion, and a pipeline that passes it can still fail across a sequence.

Teams operating in that second category can apply the protocol in Section 6 to their own pipeline without adopting any part of Pulse AI. That is the intended use: a standard a buyer can hold a vendor to, including this one.

10. Ethics and Consent

All sender identity assets used in this work were provided with explicit written consent for synthetic video generation and for use in evaluation. No likenesses of third parties, public figures, or non-consenting individuals were generated or evaluated. Negative-control comparisons draw on published face datasets [6, 7] rather than on additional consented identities.

Identity-preserving synthetic video carries obvious potential for impersonation. This framework measures whether a sender remains themselves; it

does not authenticate that the sender authorised the message, and it should not be read as a provenance or consent-verification mechanism. The intended deployment constraint is that senders generate video only of themselves, and the metric functions as a quality gate rather than a trust credential. Nothing in this protocol should be applied to generate video of a person who has not consented.

Funding

No external funding was received for this work. The author is the founder of Pulse AI, the commercial system this framework evaluates; this constitutes a competing interest and is disclosed accordingly.

References

[1] Deng, J., Guo, J., Xue, N., & Zafeiriou, S. (2019). ArcFace: Additive Angular Margin Loss for Deep Face Recognition. *CVPR*, 4690–4699.
[2] InsightFace: 2D and 3D Face Analysis Project. *github.com/deepinsight/insightface*. Model pack: buffalo_l.
[3] Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., & Sebe, N. (2019). First Order Motion Model for Image Animation. *NeurIPS*, 7135–7145.
[4] Prajwal, K. R., Mukhopadhyay, R., Namboodiri, V. P., & Jawahar, C. V. (2020). A Lip Sync Expert Is All You Need for Speech to Lip Generation In The Wild. *Proceedings of the 28th ACM International Conference on Multimedia*, 484–492. doi:10.1145/3394171.3413532
[5] Rössler, A., Cozzolino, D., Verdoliva, L., Riess, C., Thies, J., & Nießner, M. (2019). FaceForensics++: Learning to Detect Manipulated Facial Images. *ICCV*, 1–11. doi:10.1109/ICCV.2019.00009
[6] Huang, G. B., Ramesh, M., Berg, T., & Learned-Miller, E. (2007). Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. *University of Massachusetts, Amherst, Technical Report 07-49*, October 2007.
[7] Karras, T., Laine, S., & Aila, T. (2019). A Style-Based Generator Architecture for Generative Adversarial Networks. *CVPR*, 4401–4410. (FFHQ dataset.)

Citation

Das, A. (2026). *Identity Persistence in AI-Generated Video: A Sequence Evaluation Framework for Founder-Led Outbound.* Pulse AI Working Paper. Zenodo. https://doi.org/10.5281/zenodo.21620531