Node-Level Identity Attention for Multi-Face Frames

A Controlled Landmark-Graph Study

Research article • Muhammad Usman • Superior University Lahore • 14 October 2026

Abstract

When an AI video generator puts two or more faces in the same frame, their identities tend to leak into each other — faces blend, swap, or fade. Most existing fixes work with coarse face regions or pixel masks, but published evidence shows these masks cannot fully stop the leakage. This study asks a simple question: what happens if we route identity through anatomically indexed facial landmarks instead of coarse regions?

Using 468 MediaPipe landmarks per face, we build a nearest-landmark ownership map and test it in 200 controlled trials per condition. Region-only routing managed just 0.620 association accuracy with a leakage score of 0.380. Node routing with cross-attention isolation reached 1.000 association and a mean wrong-identity margin of −0.401; closing the self-attention and convolutional paths pushed the margin to −0.640 and −0.829. The results are controlled simulations, not real-world recognition claims — but they motivate the next step: real face images, and then integration into a diffusion backbone.

Keywords: multi-face identity, facial landmarks, node routing, attention isolation, identity leakage, diffusion video generation

> *Study design note: the experiments in this article use controlled score simulations and synthetic landmark geometry to isolate the routing mechanism. The results are method-development evidence, not a claim of real-world identity-recognition performance.*

1. Research Question and Contributions

The core question: can an anatomically indexed node-ownership signal route identity more precisely — and more auditably — than a coarse face-region mask? Our working hypothesis is that identity information travels along three paths inside a generator, and all three must be controlled: (1) injection through cross-attention, (2) propagation through self-attention, and (3) spatial blurring through convolutional feature mixing.

What this study delivers:

- A working pipeline that extracts 468 MediaPipe landmarks per face and normalizes them into a canonical form.
- A node-derived ownership map designed to plug into block-diagonal attention masks later.
- An ID-Patch-style association metric plus a continuous wrong-identity margin that catches near-confusions a binary score would miss.
- A fixed ablation protocol with machine-readable results (JSON/CSV) and visual diagnostics.
- A clear route from this landmark evaluation to full diffusion integration.

2. Method

2.1 Landmark graph and node descriptors

For each detected face, the MediaPipe Face Landmarker gives us 468 normalized 3D landmarks. Each node's descriptor combines two things: the landmark geometry (Procrustes-normalized 2D positions, with translation, scale, and principal-axis rotation factored out) and a small local grayscale patch sampled around the landmark. This is deliberately a proxy — good enough for routing experiments, while a proper face-recognition encoder is reserved for the real-image stage.

2.2 Ownership routing

Given two face graphs, every pixel (or latent position) is assigned to its nearest visible landmark, producing a Voronoi-style identity partition — the owner map. For diffusion, the interface is simple: owner[p] ∈ {identity 0, identity 1, ..., background}, plus a soft node-affinity vector and per-identity node features. In the model itself, the owner map becomes additive $-\infty$ logits on every cross-identity attention entry.

Figure 1. Landmark ownership partition for two synthetic 468-node face graphs, illustrating the routing geometry used in the experiment.

2.3 Metrics

Let S_{ij} be the similarity between observed face i and reference identity j. The predicted identity is $\text{argmax}_j S_{ij}$. Association accuracy is the fraction of faces matched to the correct reference; the discrete leakage score is L = 1 − association accuracy. We also report a continuous wrong-identity margin:

$$m_i = \max_{j \neq i} S_{ij} - S_{ii} \qquad \bar{m} = \text{mean}_i(m_i)$$

A negative margin means the correct reference beats every wrong one; a margin near zero means the decision was fragile even when the argmax happened to be right. That distinction matters — a binary score can hide near-confusions.

3. Controlled Experiments

Fixed seed 20261014, two identities, 200 trials per condition, Gaussian perturbations around prescribed correct/incorrect similarity means. Four conditions:

1. Region-only: correct and wrong scores nearly tied — ambiguous coarse routing.
2. Node + cross: node routing with identity-specific injection, propagation left open.
3. Node + cross + self: cross-identity self-attention additionally isolated.
4. Node + all three paths: grouped/gated convolutional isolation additionally modeled.

3.1 Results

Table 1. Controlled ablation results. The score distributions are specified to isolate routing strength, so this evaluates the metric and implementation logic — not a trained diffusion model.

Condition	Association	Leakage L	Wrong-ID margin
region only	0.620	0.380	−0.029
node + cross	1.000	0.000	−0.401
node + cross + self	1.000	0.000	−0.640
node + all three paths	1.000	0.000	−0.829

Figure 2. Controlled ablation results (red series = 1 + margin; higher means larger separation from the nearest wrong identity). Figure 3. Similarity matrix from a final-path control trial — diagonal dominance is the desired pattern.

3.2 Implementation checks

The implementation ships with tests for perfect association, swapped identities, and ownership-map coverage. Every run emits JSON, CSV, PNG figures, and a text log; Figure 4 shows the diagnostic log with the fixed seed and aggregate outputs behind Table 1. Three unit tests passed, Python syntax compiled cleanly, the MediaPipe model asset was present, and all figures and result files generated successfully.

4. Discussion

4.1 Interpretation

Three takeaways. First, the association metric cleanly separates correct from swapped identities. Second, the continuous margin separates robust matches from near-ties — information the binary score throws away. Third, the ownership map is directly compatible with cross-attention and self-attention masking, which is exactly the interface the diffusion stage needs.

4.2 Limitations

- Real-person identity preservation still needs testing on a consented image set.
- The current descriptor is a geometry-and-patch proxy; ArcFace/CurricularFace evaluation belongs to the next study.
- This work comes before UNet, DiT, and video-diffusion integration.

- The grouped-convolution effect must be measured inside the actual diffusion architecture.
- Fair comparison with published baselines needs a shared real-image benchmark and confidence intervals.

4.3 The main weakness in the current metric

The proxy signature leans on landmark geometry plus grayscale patches. Geometry correlates with identity only weakly and shifts with expression and pose; grayscale patches are lighting-sensitive. The next version must evaluate with a recognized face embedding from a held-out evaluator, keeping the routing representation separate to avoid metric leakage.

5. Phase 3 Protocol: Real-Face and Diffusion Validation

5. Collect two consenting reference identities plus held-out images/video frames, recording pose, expression, lighting, distance, and occlusion metadata.
6. Swap the proxy signature for ArcFace on the representation side and CurricularFace/FaceNet-style similarity on the evaluation side.
7. Build a two-face frame benchmark with known face-to-reference labels and a face detector/landmark tracker.
8. Run the fixed ablation: cross-attention mask only → + block-diagonal self-attention → + grouped/gated convolution.
9. Report association accuracy, leakage score, continuous margin, temporal drift, merge events, runtime, and confidence intervals.
10. Only after the image benchmark passes, integrate the owner map into a diffusion backbone and repeat the same protocol.

6. Reproducibility Package

The accompanying package contains source code, requirements, the MediaPipe model asset, unit tests, the experiment generator, raw JSON/CSV outputs, the diagnostic log, and figures. To reproduce:

```
cd node_identity_attention_prototype
python3 phase2_experiment.py
python3 -m pytest -q
```

Output is deterministic under the fixed random seed. Real-image runs must be stored separately with consent and provenance metadata.

References

1. Zhang, Y. et al. ID-Patch: Robust ID Association for Group Photo Personalization. CVPR 2025. https://arxiv.org/abs/2411.13632

2. Kim, H. et al. InstantFamily: Masked Attention for Zero-shot Multi-ID Image Generation. arXiv:2404.19427. https://arxiv.org/abs/2404.19427

3. Wang, Q. et al. InstantID: Zero-shot Identity-Preserving Generation in Seconds. arXiv:2401.07519. https://arxiv.org/abs/2401.07519

4. Papadopoulos, A. et al. Face-GCN: A Graph Convolutional Network for 3D Face Recognition. arXiv:2104.09145. https://arxiv.org/pdf/2104.09145

5. Google MediaPipe. Face Landmarker / Face Mesh model documentation. https://ai.google.dev/edge/mediapipe/solutions/vision/face_landmarker

6. Usman, M. Node Identity Attention Research Report. 14 October 2026. Source document for the present study.