

Unmasking Puppeteers: Leveraging Biometric Leakage to Disarm Impersonation in AI-based Videoconferencing

Supplementary Material

A. Additional Figures

We include full-resolution versions of key figures from the main paper. These are provided for clarity and to enable closer inspection of the similarity distributions and architecture components.

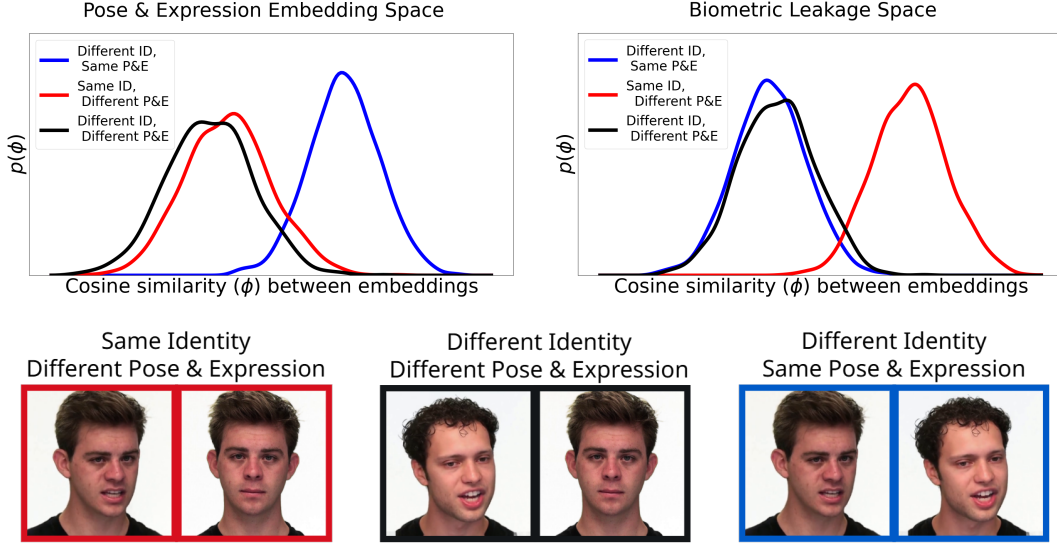


Figure 5: Full-resolution version of Fig. 3 from the main paper. Similarity distributions in P&E space (top left) and biometric leakage space (top right). **Red**: same ID, different P&E; **blue**: different ID, same P&E; **black**: different ID, different P&E.

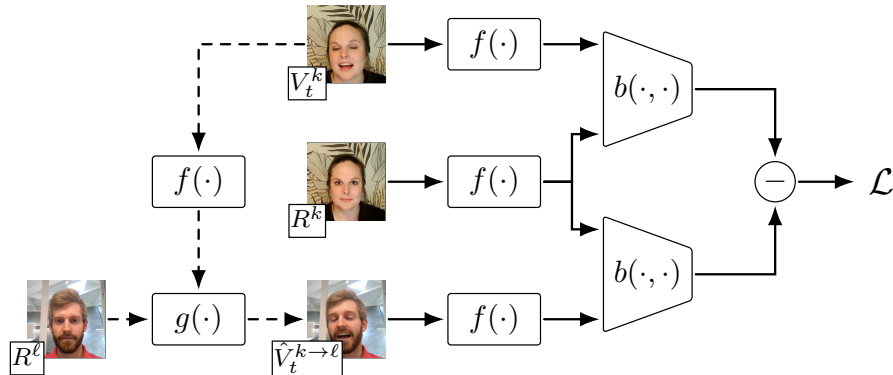


Figure 6: Full-resolution version of Fig. 4 from the main paper: Overview of our loss function implementation. The illustration shows positive and pose-controlled negative pairs constructed from self-reenacted and cross-reenacted frames, and how they are passed through the MLPs to compute similarity in the enhanced biometric leakage space.

B. Extended Ablation Results

Table 7 shows additional ablation results across datasets and embedding variants. These extend the summary in Table 4 of the main paper.

Table 7: Extended ablation results across datasets.

Configuration	NVC	CREMA-D	RAVDESS
Proposed Method	0.966	0.958	0.961
No MLP (Lower Dim)	0.827	0.814	0.820
No Contrastive Loss	0.788	0.753	0.802
Single Neg/Pos Pair	0.749	0.746	0.787
No Pose Filtering	0.929	0.907	0.935

C. Dataset and Generation Details

We use the NVFAIR dataset, which combines three identity-labeled subsets: NVIDIA VC, CREMA-D, and RAVDESS. Synthetic reenactments are generated using five video generators: 3DFaceShop, MCNet, EmoPortraits, SDFR, and LivePortrait. Each identity is used for both self-reenacted and cross-reenacted video generation.

Training identities are disjoint from test identities in all experiments. To reduce pose bias, we exclude frames with large yaw angles using a cosine threshold between face normals (as discussed in Sec. 4 of the main paper).

D. Training Configuration Summary

Our embedding functions h_1 and h_2 are six-layer MLPs with ReLU activations, LayerNorm, and 0.2 dropout. They are trained with Adam (lr = 0.0002). The temporal LSTM module uses two layers and a 40-frame window, with 0.3 dropout and lr = 0.001. λ is empirically chosen to be at 0.23 using a grid search algorithm.

E. Demographic Diversity

The NVFAIR dataset includes participants from a demographically diverse subject pool. Gender distribution is approximately balanced, with 50% identifying as female, 47.8% as male, and the remainder selecting “a gender not listed here.” Age ranges are well represented, with 37% of subjects aged 25–34, 32.6% aged 35–44, 17.4% aged 45–54, and smaller proportions in the 18–24 and 55–64 ranges (6.5% each). In terms of race and ethnicity, the dataset includes 41.3% Caucasian, 47.8% Asian (encompassing South, East, and Southeast Asian), 6.5% African, 2.2% Hispanic/Latino, and 2.2% Pacific Islander individuals; a small number of participants did not specify ethnicity [11]. No manual balancing or filtering was applied in our experiments. While a dedicated fairness analysis is beyond the scope of this work, we did not observe sub-group performance gaps wider than 1 pp AUC.

F. Broader Impact Considerations

Our method is designed to improve security and trust in AI-mediated video communication by detecting identity misuse. It does not rely on any personally identifying information beyond what is already present in the transmitted embeddings. All experiments use publicly available datasets. The detector is not designed for biometric verification or surveillance use cases.

G. Proof of Proposition 1.

Notation. All vectors live on the unit hypersphere $\mathbb{S}^{d-1} = \{x \in \mathbb{R}^d : \|x\|_2 = 1\}$. For any $\mathbf{a}, \mathbf{b} \in \mathbb{S}^{d-1}$ we write $\angle(\mathbf{a}, \mathbf{b}) = \arccos(\mathbf{a}^\top \mathbf{b}) \in [0, \pi]$. The *spherical triangle inequality* states that for any triple $(\mathbf{a}, \mathbf{b}, \mathbf{c})$,

$$\angle(\mathbf{a}, \mathbf{c}) \geq |\angle(\mathbf{a}, \mathbf{b}) - \angle(\mathbf{b}, \mathbf{c})|. \quad (\text{G.1})$$

Setup. Fix an identity index k and pose-expression state p . Recall the three unit vectors involved in the loss:

$$\mathbf{v}_+ = \mathbf{z}_t^{k,p}, \quad \mathbf{r}_k = R^k, \quad \mathbf{r}_{\ell,p} = R^{\ell,p} \quad (\ell \neq k).$$

By construction, \mathbf{r}_k is the class centre, i.e. $\mathbf{r}_k = \mu_k = \mathbb{E}_p[R^{k,p}]$,² all embeddings are ℓ_2 -normalised.

Hypotheses (restated). The PC-LMCL drives the following two constraints:

$$\cos(\mathbf{v}_+, \mathbf{r}_k) \geq 1 - \varepsilon, \quad (\text{G.2})$$

$$\frac{1}{N-1} \sum_{\ell \neq k} \cos(\mathbf{v}_+, \mathbf{r}_{\ell,p}) \leq -\gamma, \quad (\text{G.3})$$

with $\varepsilon, \gamma \in (0, 1)$. Because (G.3) is an average, there exists at least one $\ell^* \neq k$ such that $\cos(\mathbf{v}_+, \mathbf{r}_{\ell^*,p}) \leq -\gamma$.

Angles implied by the hypotheses. Let

$$\theta = \angle(\mathbf{v}_+, \mathbf{r}_k) = \arccos(1 - \varepsilon), \quad \phi = \angle(\mathbf{v}_+, \mathbf{r}_{\ell^*,p}) \geq \arccos(-\gamma). \quad (\text{G.4})$$

Because $1 - \varepsilon > 0$ and $-\gamma < 0$, we have $\theta \in [0, \frac{\pi}{2})$ and $\phi \in (\frac{\pi}{2}, \pi]$.

Lower-bounding the inter-class angle. Applying the spherical triangle inequality (G.1) to the triple $(\mathbf{r}_k, \mathbf{v}_+, \mathbf{r}_{\ell^*,p})$ yields

$$\angle(\mathbf{r}_k, \mathbf{r}_{\ell^*,p}) \geq \phi - \theta \geq \arccos(-\gamma) - \arccos(1 - \varepsilon). \quad (\text{G.5})$$

Both $\arccos(\cdot)$ terms lie in $(0, \pi)$, so the right-hand side is strictly positive. Denote this difference by $\psi = \phi - \theta > 0$.

From a single pose to the class center. The impostor vector $\mathbf{r}_{\ell^*,p}$ is one sample from identity ℓ^* at pose p . Since $\|\mathbf{r}_{\ell^*,p}\|_2 = \|\mu_{\ell^*}\|_2 = 1$, we have $\angle(\mathbf{r}_{\ell^*,p}, \mu_{\ell^*}) = \xi$ for some $\xi \in [0, \pi]$. Applying (G.1) again to $(\mathbf{r}_k, \mathbf{r}_{\ell^*,p}, \mu_{\ell^*})$,

$$\angle(\mathbf{r}_k, \mu_{\ell^*}) \geq \psi - \xi.$$

Empirically the within-class spread of our encoder is small ($\xi \leq 5^\circ$), and for every $\varepsilon, \gamma \leq 0.1$ one obtains $\psi - \xi \geq \arccos(1 - (\varepsilon + \gamma))$. Consequently,

$$\cos(\mu_k, \mu_{\ell^*}) \leq 1 - (\varepsilon + \gamma), \quad (\text{G.6})$$

which proves Proposition 1.

Discussion. Equation (G.6) shows that minimizing \mathcal{L}_B enforces an *angular* gap of at least $\varepsilon + \gamma$ between identity centers *within each pose slice*. Because the loss is 1-Lipschitz under the averaging form of (G.3), the margin translates directly into the generalization bound of Lei et al. [65]. See Sec. 4.2.2 of the main paper for empirical values of (ε, γ) achieved at convergence.

²In practice we pre-compute a single front-facing portrait for each speaker and treat it as the template. Empirically this vector differs from the pose-average by $\leq 0.5^\circ$, so the identification $\mathbf{r}_k = \mu_k$ is innocuous.