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

Danial Samadi Vahdati[1], Tai Duc Nguyen[1], Koki Nagano[2], David Luebke[2], Orazio Gallo[2], Ekta Prashnani[2], Matthew Stamm[1]
[1]Drexel University [2]NVIDIA
Multimedia & Information Security Lab

Key Insight

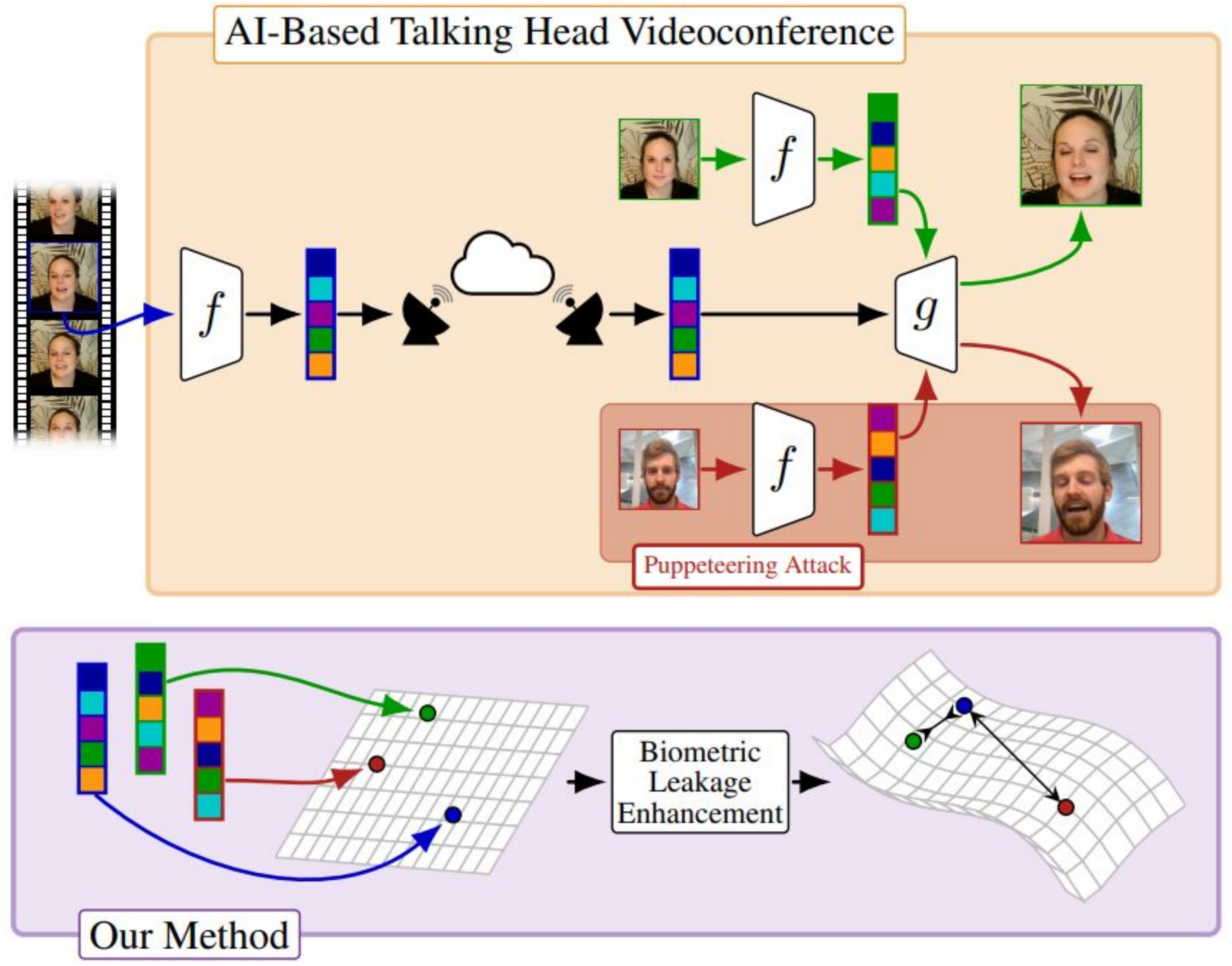
Biometric Leakage in Latent Space

- Pose/expression embeddings inherently leak identity cues
- Physical anatomy (jaw, eye spacing, lips) tied to motion
- Same pose from different identities produces distinct embeddings
- We exploit this leakage for defense

Problem

The Puppeteering Threat

- Attacker swaps victim's identity at call initialization
- Sender transmits unauthorized reference portrait
- Attacker's P&E drives victim's face
- Receiver sees video of wrong person
- Traditional deepfake detectors fail



Novelty

Enhanced Biometric Leakage (EBL) Space

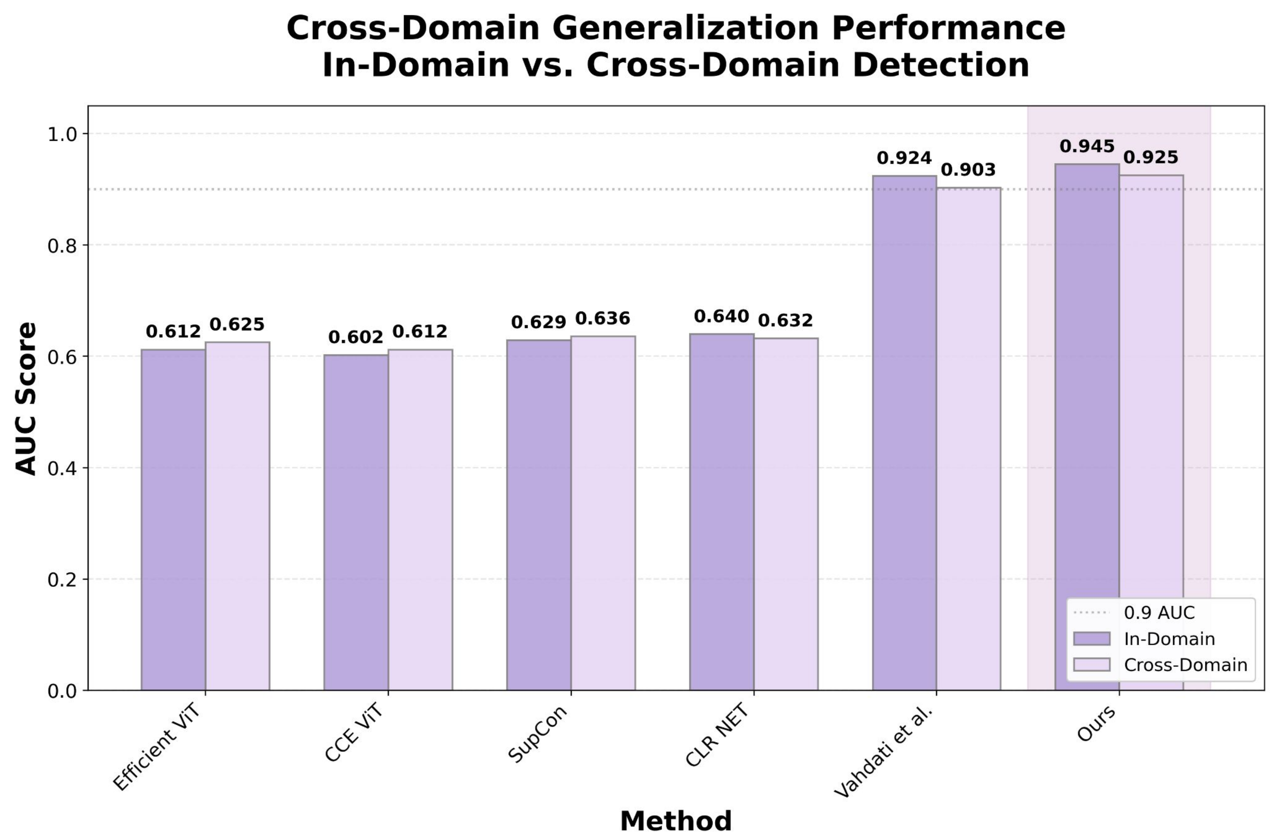
- Re-encodes embeddings to amplify identity, suppress pose
- Pose-Conditioned Contrastive Loss:
  - Pulls same identity together across poses
  - Pushes different identities apart
- LSTM aggregates 40 frames for stability

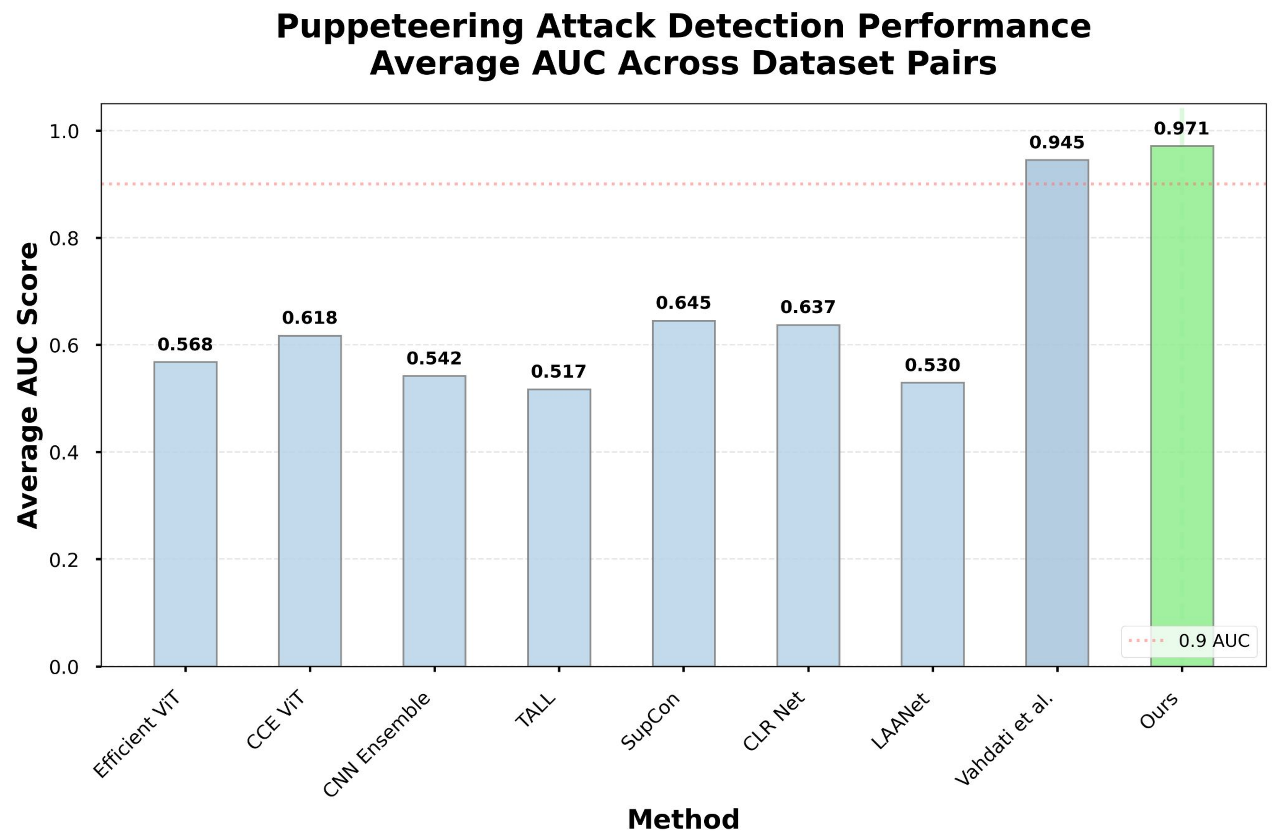
Solution

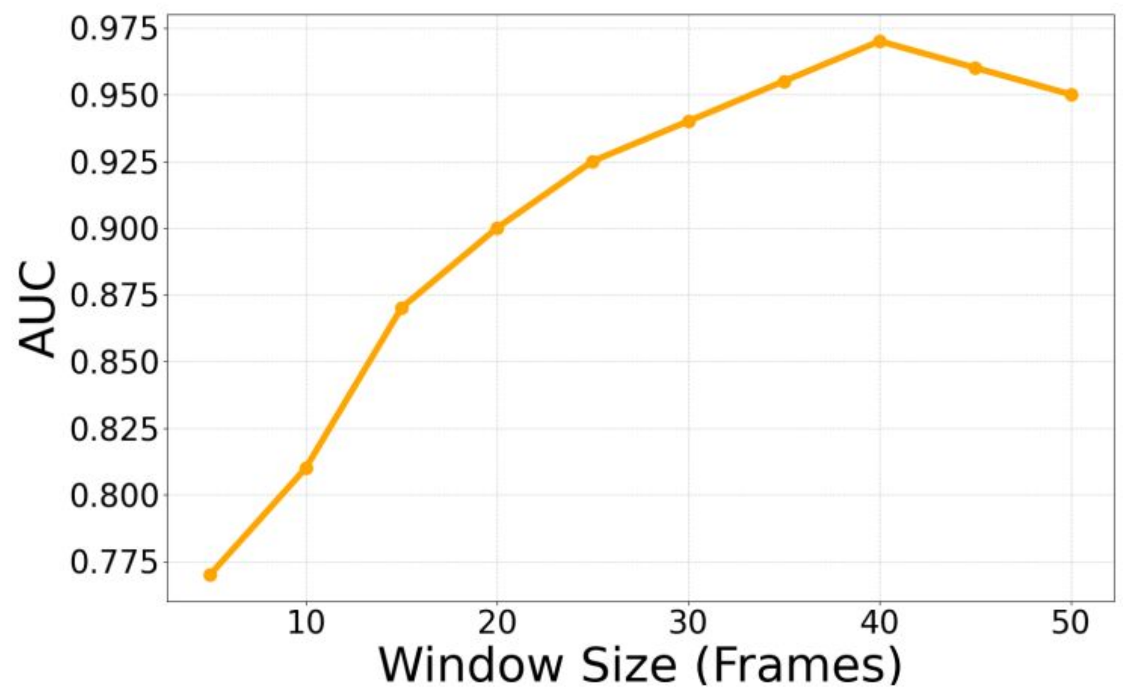
Our Defense: Latent-Space Authentication

- Operates entirely in latent space (no RGB reconstruction)
- Compares driving vs. target identity embeddings
- Real-time detection: 75 FPS, <1M parameters
- No enrollment or landmarks required

Results and Performance







Key Result

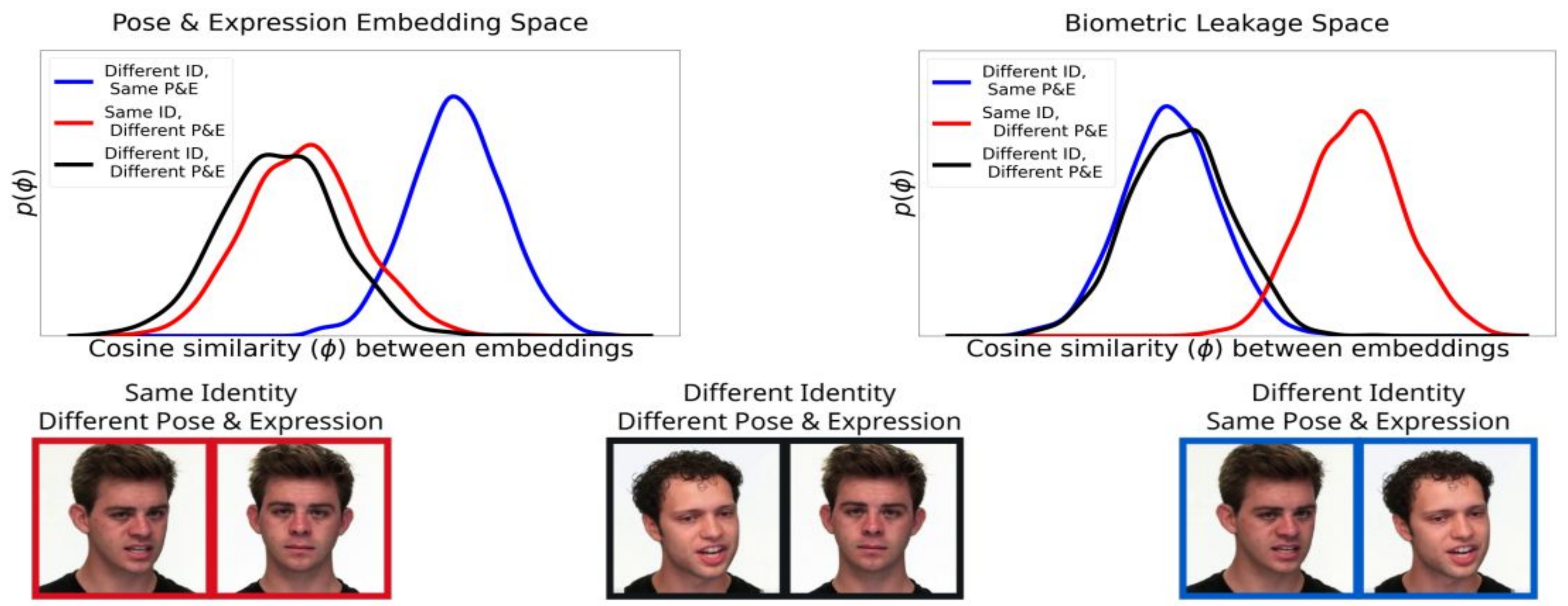
- Trained on NVFAIR (NVIDIA-C subset), tested on RAVDESS & CREMA-D
- 94.5% AUC in-domain vs. 92.5% cross-domain
- Only 2% performance drop (vs. 5-10% for baselines)
- Outperforms all methods in both settings
- Deepfake detectors fail completely cross-domain

Key Result

- Achieves 97.7% average AUC across all tests
- Consistent performance on 5 different generators
- Tested on 3 datasets (NVIDIA-VC, RAVDESS, CREMA-D)
- Outperforms all deepfake detectors (>30% gap)
- Minimum AUC above 95% across all scenarios

Temporal Fusion

- AUC improves from 77% to 97% with temporal fusion
- Optimal window size: 40 frames (~1.3s at 30fps)
- LSTM aggregates frame-level similarity scores
- Performance plateaus beyond 40 frames
- Temporal context crucial for stable detection



How It Works

Core Innovation

Why EBL Space works

- P&E space: pose variation masks identity signal
- EBL space: identity signal dominates pose variation
- Contrastive loss pulls same identity together
- Hard negatives (same pose, diff ID) push identities apart
- Result: reliable identity discrimination in real-time

Our Method

1. Enhanced Biometric Leakage (EBL) Space:

$$b(z_t, R) = \cos(h_1(z_t), h_2(f(R)))$$

2. Pose-Conditioned Large-Margin Cosine Loss:

$$\mathcal{L}_P = 1 - b(z_t^{k,p}, R^k)$$

$$\mathcal{L}_P = 1 - b(z_t^{k,p}, R^k)\mathcal{L}_N = \frac{1}{N-1}\sum_{\ell \neq k} b(z_t^{k,p}, R^{\ell,p})$$

$$\mathcal{L}_B = \mathcal{L}_P + \lambda \mathcal{L}_N$$

3. LSTM Temporal Fusion:

$$\phi = \{\phi_1, \phi_2, \dots, \phi_W\}$$

$$y = \mathrm{LSTM}(\phi_1, \dots, \phi_W)$$