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

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

Abstract

AI-based talking-head videoconferencing systems reduce bandwidth by sending a compact pose-expression latent and re-synthesizing RGB at the receiver—but this latent can be "puppeteered," letting an attacker hijack a victim's likeness in real time. Because every frame is synthetic, deepfake and synthetic video detectors fail outright. To address this security problem, we exploit a key observation: the pose expression latent inherently contain biometric information of the driving identity. Therefore, we introduce the first *biometric leakage defense* without ever looking at the reconstructed RGB video: a pose-conditioned, large-margin contrastive encoder that isolates persistent identity cues inside the transmitted latent while cancelling transient pose and expression. A simple cosine test on this disentangled embedding flags illicit identity swaps as the video is rendered. Our experiments on multiple talking-head generation models show that our method consistently outperforms existing puppeteering defenses, operates in real-time, and shows strong generalization to out-of-distribution scenarios.Our code and trained models are available at https://github.com/MISLresearch/Unmasking-Puppeteers-Neurips25.

1 Introduction

Advances in generative AI have enabled the creation of hyper-realistic synthetic videos. This has led to the development of many new technologies, including both avatar-based communication systems [1–3] and AI-based videoconferencing systems [4–7]. AI-based talking head videoconferencing systems are receiving increasing attention due to their significant bandwidth savings [8]. Instead of continuously encoding and transmitting each video frame of a speaker, these systems only transmit embeddings that capture a speaker's pose and facial expression. Then, a generative AI system at the receiver's end uses these embeddings, in conjunction with an initial representation of the speaker, to create an accurately reconstructed video.

Unfortunately, AI-based talking head videoconferencing systems [5, 4, 8, 7, 9] are vulnerable to a new form of information attack known as "puppeteering" [10]. In this attack, a malicious user at the sender side transmits an unauthorized representation of a different target speaker when a video call is initiated. As a result, the identity in the video that the receiver constructs is different than the identity of the person who actually controls the video [10, 11].

While many forensic approaches, such as deepfake [12–17] and synthetic video detectors [18–21], aim to expose unauthorized AI-generated media, they operate by identifying evidence that a video has been synthetically generated. However, in AI-based talking-head videoconferencing, *every video is AI-generated*, rendering existing media forensic approaches unable to detect puppeteering.

[*]Corresponding author: ds3729@drexel.edu

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

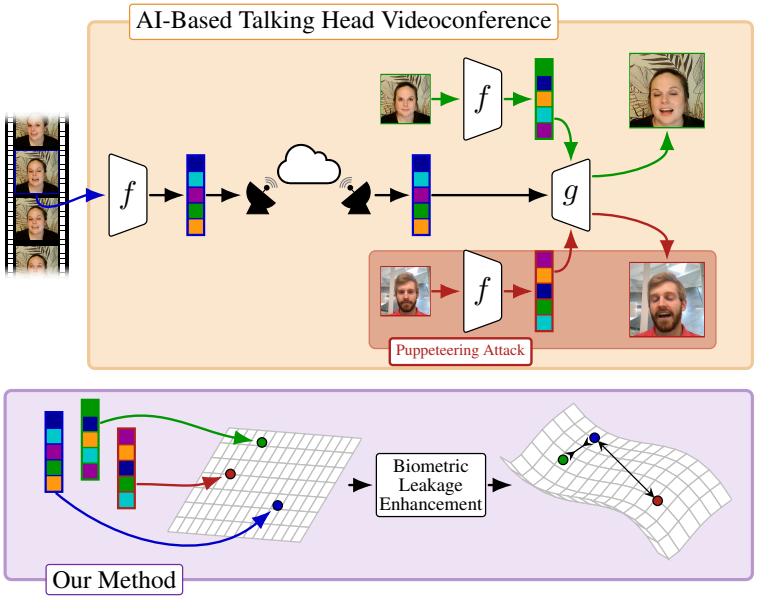

Figure 1: AI-based talking-head generators transmit only a compact pose-and-expression embedding for low-bandwidth videoconferencing, but remain vulnerable to puppeteering attacks that swap in a different identity for live impersonation. Our defense capitalizes on biometric signals inadvertently leaked in these embeddings to reveal mismatches between the driving speaker and the reconstructed identity in real time.

Puppeteering attacks pose a fundamentally different threat from those typically addressed by media forensics. Instead of asking "Is this video real or AI-generated?" one key question must be "Is this AI-generated video being driven by an authorized identity, or has someone hijacked it?" As talking-head videoconferencing technology continues to advance, it is essential to mitigate puppeteering attacks before these systems are widely deployed. Addressing these threats proactively will help ensure secure and trustworthy communication over these systems.

Few defenses address puppeteering. Prashnani et al. introduced Avatar Fingerprinting [11], which works by enrolling each user's characteristic facial motion patterns into the detector's training data. Although this could help, it requires significant user-specific data to build reliable motion signatures. In contrast, Vahdati et al. avoid enrollment by decoding and re-encoding the received video for comparison [10]; the round-trip is compute-heavy and adds distortions that hurt accuracy.

In this paper we introduce a real-time, enrollment-free defense that detects puppeteering attacks entirely in the latent domain. Our key insight is that the pose-and-expression embeddings already transmitted by modern talking-head systems leak subtle, but reproducible, biometric signatures [22, 23]. We learn a compact Enhanced Biometric-Leakage (EBL) space in which identity cues are amplified while pose and expression variance is actively suppressed. A pose-conditioned contrastive loss drives this separation, and a lightweight temporal LSTM aggregates evidence to yield stable, millisecond-level decisions. By comparing the sender's latent identity with the target identity rendered at the receiver, our system flags any mismatch without decoding RGB frames, tracking landmarks, or collecting per-user motion profiles. Extensive experiments across many generators and datasets show that our method achieves state-of-the-art detection performance while working in real-time.

In summary, our contributions are as follows.

- We present the first method that operates solely on data already available at the receiver, requiring no user enrollment or additional sensors.
- We formalize and harness identity leakage inherent in pose–expression embeddings, turning a previously ignored vulnerability into a defensive signal.
- We develop a novel loss and training protocol to learn a low-dimensional EBL representation that maximizes identity separability while nullifying pose and expression.

Figure 2: Illustration of three datasets (NVIDIA-VC [11], RAVDESS [54], CREMA-D [55]) shown across three columns each. Row 2 indicates the type of frame displayed in Row 3: Reference, Self-Reenacted, or Cross-Reenacted.

- Though extensive experiments across fifteen generator/dataset combinations, our approach achieves state-of-the-art puppeteering detection in real time, demonstrating practical viability for deployment in bandwidth-constrained videoconferencing systems.

2 Background and Related Work

Talking-Head Video Systems. Talking-head systems synthesize facial motion and speech in real time by transmitting low-dimensional embeddings of pose and expression rather than full video frames [24–27]. A generator at the receiver reconstructs a realistic face. Methods range from compressed embeddings to landmark-based updates per frame [8, 28–33], reducing bandwidth while maintaining high visual fidelity [34–40]. Systems like NVIDIA's Vid2Vid Cameo [8] and Google's Project Starline [41] demonstrate real-time, 3D teleconferencing.

Forensics for Synthetic Media. Synthetic media forensics traditionally aims to distinguish real from AI-generated content. Deepfake detectors [12–17, 42] leverage semantic cues [43–46], lighting inconsistencies [47–50], and facial texture artifacts [51–53], while synthetic video detectors [18–21] focus on identifying AI-generated traces. However, in low-bandwidth talking-head systems, every video is synthetic [10, 11], rendering these detectors ineffective - they cannot determine whether the video represents the correct identity or has been hijacked.

Puppeteering Attacks and Existing Defenses. Puppeteering attacks replace the target's appearance with the adversary's pose-and-expression vectors at call initiation, causing the receiver to synthesize a video of the wrong identity. Avatar Fingerprinting [11] mitigates this threat by first enrolling each user's facial-motion signatures; at inference it flags a session whenever the incoming motion deviates from the stored profile. Vahdati et al. [10] avoid enrollment by re-encoding the synthesized video and comparing embeddings, but their pipeline depends on facial-landmark estimation, which degrades under large head rotations and lighting artifacts introduced by video reconstruction.

In contrast, our method exploits biometric leakage *already present* in the pose-and-expression latents, removing both the enrollment phase and any reliance on facial landmarks. A contrastive objective disentangles identity from pose, yielding a compact, latency-friendly representation that generalizes to unseen generators and real-world conferencing conditions.

3 Problem Formulation

3.1 Talking Head Video Systems

AI-enabled talking-head videoconferencing systems have been proposed to minimize transmitted data by encoding and transmitting only pose and expression information from a speaker rather than an entire video frame [8, 56–62]. As this technology rapidly matures, our goal here is to develop security measures that improve the trustworthiness of these systems when they are deployed.

A videoconferencing call starts with each participant k sending a neutral reference portrait R^k to the receiver. Typically, R^k is an image of the speaker in a neutral pose. As the speaker continues to speak, each new video frame V_t^k capturing this speaker at the sender side is processed by an embedding function f to produce an embedding $z_t^k = f(V_t^k)$ that encodes the speaker's instantaneous pose and expression at time t. This embedding is transmitted to the receiver, as is shown in Fig. 1.

At the receiver side, a generator g is used to produce a video of the speaker on the basis of the received embedding z_t^k and the reference representation R^k such that

$$\hat{V}_t^{k \to k} = g(z_t^k, R^k). \tag{1}$$

Here, the superscript $\hat{V}_t^{k \to k}$ denotes that speaker k both *drives and appears* in the rendered video. We refer to the sender's identity as the "driving identity" and the reference portrait sent to the receiver as the "target identity" [11]. While such a talking-head video-conferencing system promises to be bandwidth-efficient, it exposes a vulnerability to puppeteering attacks when a malicious sender impersonates another individual.

3.2 Puppeteering Attacks

A puppeteering attack exploits the inherent trust in the reference representation transmitted to the receiver when a videoconferencing call begins. Here, an adversary, speaker ℓ, obtains a target speaker's representation R^k and substitutes it for their own without authorization at the start of the call [10]. As the call proceeds, the adversary's own video V_t^ℓ is used to derive the pose and expression vectors z_t^ℓ that are transmitted to the receiver. The receiver then uses the generator g alongside the unauthorized reference R^k and z_t^ℓ to reconstruct the video. As a result, the receiver is presented with a realistic-looking video of speaker k, controlled in real time by the speaker ℓ such that

$$\hat{V}_t^{k \to \ell} = g\big(f(V_t^\ell), R^k\big) = g\big(z_t^\ell, R^k\big), \tag{2}$$

where $\hat{V}_t^{k \to \ell}$ is an unauthorized impersonation of speaker ℓ driven by speaker k. To address this threat, one must come up with mechanisms to detect whether the receiver's generated video is driven by the same authorized identity, rather than an unauthorized impersonator.

3.3 Why Real-vs-Synthetic Detectors Fail

Deepfake and synthetic-video detectors flag frames whose pixel statistics reveal AI generation, implicitly assuming that "real" camera footage is the baseline. In bandwidth-efficient talking-head conferencing, however, every frame - even those from honest participants - is produced by a generator; "synthetic" is the norm, not the anomaly. The security question therefore shifts from "Was this video AI-generated?" to "Does the rendered face correspond to the person actually driving the latents?" Pixel-level detectors cannot answer that mapping, leaving puppeteering attacks undetectable.

4 Proposed Approach

In this paper, we present a real-time solution for detecting puppeteering attacks in talking-head videoconferencing, uniquely operating entirely in the latent domain without access to reconstructed RGB frames. To do this, we re-encode each pose-and-expression latent embedding into a compact Enhanced Biometric Leakage (EBL) embedding space that captures who is speaking while discarding how they move. A pose-conditioned contrastive loss learns this space by pulling together identity-consistent pairs and pushing apart identity-mismatched, pose-matched pairs. At run time, we simply compare each live EBL vector to the EBL embedding of the reference portrait sent at call setup; a sharp drop in similarity reveals that the driving and target identities have diverged, signalling puppeteering. For additional robustness and stable authentication, during training, we discard unreliable extreme-pose frames, and leverage an LSTM to fuse successive EBL scores to output stable, low-latency decisions. These innovations yield an algorithm that authenticates speakers in real-time, requires no enrollment or facial-landmark preprocessing. The following sections detail our approach and its implementation.

4.1 Biometric Entanglements in Pose & Expression Latent Space

Low-bandwidth talking-head systems encode each frame V_t as a pose-and-expression vector $z_t = f(V_t)$, which the receiver's generator turns back into pixels. Although designed for *geometry*, z_t inherit identity cues from the physical face that produced the motion. Measurements such as inter-ocular distance, jaw curvature, or lip thickness inevitably contaminate z_t because head pose and facial musculature cannot be sensed independently of the speaker's anatomy. Prior studies to disentangle identity from pose and expression also reported persistent identity traces in pose-conditioned representations [22, 23], which provides the empirical basis for our proposed defense.

Let R denote the static reference portrait of the nominal speaker transmitted at session start. Both endpoints apply the shared encoder f to compute the reference embedding $f(R)$, placing R and the

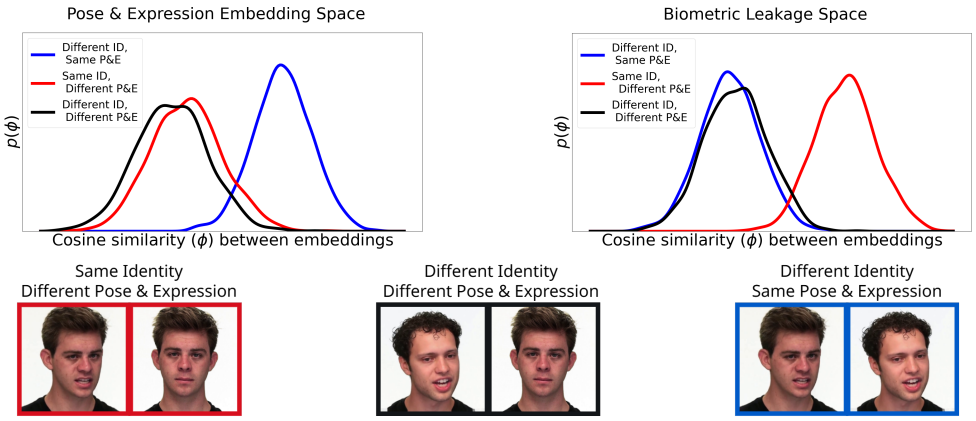

Figure 3: Similarity distributions in P&E space (left) and biometric leakage space (right). Red: same ID, diff. P&E; blue: diff. ID, same P&E; black: diff. ID, diff. P&E.

live stream $\{z_t\}$ in one latent space. A natural baseline defense is to compute the cosine similarity $s_c(z_t, f(R))$ and accept the incoming frame if the score exceeds a threshold. Unfortunately, pose and expression variability overwhelms the subtle biometric signal.

We demonstrate this effect in Fig. 3, which shows the cosine similarity distributions for three conditions: (1) same identity, differing pose/expression; (2) different identities, matched pose/expression; (3) different identities, differing pose/expression. We observe that the density for condition (2) frequently eclipses condition (1), which means that two different people exhibiting an identical yaw–pitch often *appear closer in latent space* than two frames of the same person smiling versus neutral. Raw pose-and-expression space is therefore *pose-dominated*: distance reflects "how the head moves" more than "whose head it is."

This observation is key to our puppeteering defense: **amplify latent biometric cues while suppressing pose and expression variance** so that proximity encodes identity. Simply collecting more negatives or tweaking the similarity threshold cannot meet this requirement; the representation itself must be re-shaped. In the next section, we introduce the Enhanced Biometric Leakage (EBL) space, learned directly from the latent channel, that achieves this goal without RGB reconstruction, user enrollment, or landmark tracking, enabling a real-time authentication pipeline.

4.2 Enhanced Biometric Leakage (EBL) Space

Our key observation above dictates two design imperatives: (1) craft a representation in which distances reflect identity rather than pose/expression, and (2) keep that representation compact to work in real-time. We meet both goals by re-encoding the raw pose-and-expression latent z_t into a low dimensional Enhanced Biometric Leakage vector and learning a contrastive objective that amplifies identity information while actively cancelling pose.

4.2.1 Latent re-encoding

The raw pose–expression vector z_t is dominated by geometric factors, such as yaw, pitch, mouth aperture, while biometric cues hide in low-variance directions. To amplify these cues we attach two lightweight projection heads, h_1 and h_2, that remap both the live embedding and the static reference $f(R)$ into a shared, compact metric space.

Separate heads are essential because the two inputs follow different distributions: z_t drifts frame by frame with micro-expressions, whereas $f(R)$ is fixed at call setup. Allowing independent normalization lets each head adapt to its own domain before they meet in the common space.

Each head consists of a linear layer, ReLU activation, layer normalization, and a second linear layer followed by ℓ_2 normalization, such that

$$b(z_t, R) \;=\; s_c\big(h_1(z_t),\; h_2\big(f(R)\big)\big), \tag{3}$$

where $s_c(\cdot, \cdot)$ denotes cosine similarity.

4.2.2 Pose-conditioned contrastive objective

The cornerstone of our method is a *pose-matched contrastive loss* that aligns gradient pressure with the nuisance we aim to discard. Specifically, we pursue two complementary goals inspired by large-margin hyper-spherical embedding theory [63, 64]: (1) maximize the similarity of embeddings that share the same identity even when pose/expression varies, and (2) minimize the similarity of embeddings drawn from different identities having the same pose and expression. The resulting objective comprises one term for each goal.

Positive term. Any pair that agrees in identity, regardless of pose, is treated as positive:

$$\mathcal{L}_P \;=\; 1 - b\big(z_t^{k,p},\, R^k\big), \tag{4}$$

where k indexes the speaker and p denotes the instantaneous pose-expression state. Minimizing (4) pulls together all manifestations of speaker k.

Negative term. Hard negatives are constructed by *replicating pose but swapping identity*. Concretely, we synthesise $R^{\ell,p}$, a portrait of impostor $\ell \neq k$ rendered at the identical pose p, and impose

$$\mathcal{L}_N \;=\; \frac{1}{N-1} \sum_{\ell \neq k} b\big(z_t^{k,p},\, R^{\ell,p}\big). \tag{5}$$

Because every negative matches pose, the gradient direction is orthogonal to pose variation, forcing the network to discriminate on biometric cues alone.

Total loss. The pose-conditioned Large-Margin Cosine Loss (PC-LMCL) is

$$\mathcal{L}_B \;=\; \mathcal{L}_P + \lambda\, \mathcal{L}_N, \tag{6}$$

with a single hyper-parameter λ controlling repulsion strength. With $\lambda \leq 1$, $\mathcal{L}_B$ is $(1+\lambda)$-Lipschitz (≤ 2), placing them within the margin-risk framework of Lei et al. [65].

Margin guarantee. We formalize the geometric effect of (6) as follows. We note that here, we treat each speaker's front-facing reference portrait R^k as their pose-averaged center $||R^k - \mu_k||_2 \leq 0.5°$.

Proposition 1 *Assume all embeddings are ℓ_2-normalized. If, for some $\epsilon, \gamma > 0$,*

$$\cos\big(z_t^{k,p}, R^k\big) \;\geq\; 1 - \epsilon, \quad \frac{1}{N-1} \sum_{\ell \neq k} \cos\big(z_t^{k,p}, R^{\ell,p}\big) \;\leq\; -\gamma, \tag{7}$$

then the class centers $\mu_k = \mathbb{E}_p[R^{k,p}]$ satisfy

$$\cos\big(\mu_k, \mu_\ell\big) \;\leq\; 1 - (\epsilon + \gamma), \qquad k \neq \ell. \tag{8}$$

Thus $\mathcal{L}_B$ enforces an inter-class angular margin *of at least $\epsilon + \gamma$ within each pose slice.*

Intuition. Driving $\mathcal{L}_P \to 0$ pins positives to the "north pole", while $\mathcal{L}_N \to 0$ pushes pose-matched negatives toward the antipode; the spherical triangle inequality then yields Prop. 1. The full proof appears in Appendix G.

Consequently, minimizing $\mathcal{L}_B$ *maximizes the angular decision margin* between identity manifolds while cancelling pose - an effect known to tighten generalization bounds in hyper-spherical embedding spaces [63, 66]. Prior methods ignore latent pose alignment and instead rely on RGB or landmarks, incurring latency and robustness penalties. By operating purely in the latent domain, PC-LMCL yields compact representations suitable for real-time deployment.

4.3 Training Protocol

The contrastive objective of the previous section is optimized on mini-batches organized as episodes. Each episode selects an anchor clip from speaker k, assembles identity-consistent positives drawn from other poses/expressions of k, and generates *pose-matched* negatives by rendering the same motion with impostor identities $\ell \neq k$. Because the shared encoder f remains frozen, only the two projection heads learn, allowing rapid re-training should a codec revision alter its latent format.

Table 1: Puppeteering attack detection performance measured in AUC across different dataset-generation method pairs. NVC=NVIDIA VC [11], RAV=RAVDESS [54], and CRD=CREMA-D [55]

Method	**3DFaceShop** [67]			**MCNET** [68]			**EMOPortraits** [69]			**SDFR** [70]			**LivePortrait** [71]			**Avg.**		
	NVC	**RAV**	**CRD**	**NVC**	**RAV**	**CRD**	**NVC**	**RAV**	**CRD**	**NVC**	**RAV**	**CRD**	**NVC**	**RAV**	**CRD**	**NVC**	**RAV**	**CRD**
Efficient ViT [12]	.575	.580	.562	.573	.573	.581	.560	.547	.580	.569	.579	.595	.578	.578	.591	.571	.551	.582
CCE ViT [12]	.632	.608	.618	.628	.618	.616	.624	.630	.607	.627	.622	.606	.639	.615	.575	.630	.619	.604
CNN Ensemble [13]	.540	.559	.550	.516	.536	.554	.535	.520	.537	.518	.570	.561	.527	.561	.546	.527	.549	.550
TALL [14]	.535	.511	.505	.531	.506	.504	.560	.509	.500	.546	.505	.501	.532	.501	.515	.541	.506	.505
SupCon [15]	.679	.632	.640	.651	.621	.632	.676	.649	.625	.681	.648	.619	.674	.639	.615	.672	.638	.626
CLR Net [72]	.620	.639	.635	.629	.636	.684	.634	.624	.641	.635	.634	.681	.639	.632	.600	.631	.633	.648
LAANet [16]	.538	.525	.518	.539	.519	.558	.518	.528	.556	.500	.526	.543	.513	.513	.523	.522	.527	.540
Vahdati et al. [10]	.954	**.960**	.954	.951	.950	.948	.944	.949	.952	.918	.913	.908	.959	**.966**	.958	.945	.947	.944
Ours	**.984**	.956	**.958**	**.989**	**.978**	**.984**	**.962**	**.974**	**.983**	**.970**	**.978**	**.952**	**.982**	.956	**.962**	**.977**	**.968**	**.968**

Extreme Pose Exclusion. Before each update we discard frames whose head-pose estimate is unreliable. For instance, some poses may deviate substantially from the reference image, or, the speaker's face can be substantially occluded. This extreme-pose exclusion keeps the gradient focused on informative biometric variation.

To accomplish this, we estimate face normals n_R and n_t and prune any frame where their cosine similarity $s_c(n_t, n_R)$ exceeds a threshold τ, effectively removing views with large yaw, occlusion, or visibility loss. This keeps the gradient focused on meaningful biometric variation.

4.4 Temporal Fusion

While each individual biometric similarity score provides an instantaneous assessment of the mismatch between the driving speaker and the target identity, per-frame measurements can be noisy or inconclusive under challenging poses or brief occlusions. To address this, we aggregate similarity scores over a window of W consecutive frames and feed these into an LSTM. The LSTM learns to capture temporal patterns indicative of sustained mismatches characteristic of puppeteering attacks.

Let $\phi = \{\phi_1, \phi_2, \ldots, \phi_W\}$ be the sequence of similarity scores collected over W frames. We feed $\mathbf{s}$ into the LSTM, which outputs a final score y representing the probability that the sequence is puppeteered. We train the LSTM by minimizing the binary cross-entropy loss between y and a ground-truth label t, where $t = 1$ if the video segment is puppeteered and $t = 0$ otherwise, such that

$$\mathcal{L}_{\text{LSTM}} \;=\; -t \log(y) - (1 - t) \log(1 - y). \tag{9}$$

This learnable aggregator is novel in the context of puppeteering defense: previous methods either majority-vote over heuristic landmarks or revisit pixel space, both of which are fragile under compression artifacts. As the evaluation will show, the combination of EBL embedding and temporal fusion sets a new state of the art in puppeteering detection without sacrificing latency.

5 Experiments and Results

5.1 Experimental Setup

Datasets. We conduct our experiments using the NVFAIR [11] pooled dataset because it incorporates together a large set recorded video-conference calls in a controlled setting environment. This dataset includes three subsets: (1) NVIDIA VC [11] (natural video calls), (2) CREMA-D [55] (studio-recorded expressions), and (3) RAVDESS [54] (studio-recorded emotional speech). For each subset, we generate both self-reenacted (same identity) and cross-reenacted (puppeteered) videos using five state-of-the-art methods—MCNet [68], 3DFaceShop [67], SDFR [70], EmoPortraits [69], and LivePortrait [71], resulting in total of 15 dataset-method combinations. We note that the identities used for evaluation are strictly disjoint from those in training. Further details are provided in Tab. 3.

Metrics. We report the average detection AUC to allow direct comparisons with prior work.

Competing Methods. As puppeteering detection is a new problem, we compare against the state-of-the-art method by Vahdati et al. [10]. To highlight the distinction from deepfake detection, we also evaluate seven leading deepfake detectors: Efficient ViT [12], CCE ViT [12], CNN Ensemble [13], TALL [14], SupCon [15], CLRNet [72], and LAA-Net [16].

Table 2: Puppeteering attack detection performance measured in AUC of ours and competing methods when trained only on the NVIDIA-VC [11] subset (In-domain) and tested on the CREMA-D [55] and RAVDESS [54] subsets (Cross-domain).

Method	**3DFaceshop** [67]		**MCNET** [68]		**EMOPortraits** [69]		**SDFR** [70]		**LivePortrait** [71]		**Avg.**	
	In-domain	**Cross-domain**	**In-domain**	**Cross-domain**	**In-domain**	**Cross-domain**	**In-domain**	**Cross-domain**	**In-domain**	**Cross-domain**	**In-domain**	**Cross-domain**
Efficient ViT [12]	.592	.580	.608	.635	.631	.604	.628	.659	.599	.647	.612	.625
CCE ViT [12]	.573	.598	.577	.587	.592	.616	.629	.651	.638	.609	.602	.612
CNN Ensemble [13]	.614	.629	.606	.589	.582	.516	.600	.633	.552	.595	.591	.592
TALL [14]	.626	.639	.548	.516	.562	.508	.603	.672	.561	.571	.580	.581
SupCon [15]	.637	.659	.592	.620	.599	.628	.652	.603	.664	.671	.629	.636
CLR NET [72]	.643	.627	.649	.633	.570	.625	.594	.638	.644	.638	.640	.632
LAANet [16]	.560	.558	.528	.503	.595	.568	.535	.561	.582	.617	.560	.561
Vahdati et al. [10]	.920	.895	.942	**.921**	.927	.894	.904	.890	.929	.915	.924	.903
Ours	**.948**	**.914**	**.950**	.919	**.936**	**.917**	**.951**	**.944**	**.939**	**.931**	**.945**	**.925**

Table 3: Statistics of the datasets and generated data used in this paper.

Dataset	# of IDs	# Authorized Use Videos			# Puppeteered Videos		
		Train	Test	Total	Train	Test	Total
NVIDIA-VC (NVC) [11]	46	1,331	439	1,770	41,261	3,951	45,212
RAVDESS (RAV) [54]	24	704	264	968	10,560	1,320	11,880
CREMA-D (CRD) [55]	91	5,154	1,558	6,712	319,548	28,044	347,592

Table 4: Detection AUC of our proposed method and its alternative design choices.

Component	Method	AUC	RER%
	Proposed	**0.966**	–
Biometric Space Network Design	No MLP Lower Dim. Projection	0.827	80.35
	No Biometric Leakage Network	0.635	90.68
Training Protocol	Single Negative + Single Positive	0.749	86.45
	No Biometric Contrastive Loss	0.788	83.96
	No Extreme Pose Exclusion	0.929	52.11

We exclude Avatar Fingerprinting [11], which requires an enrollment phase where each user submits authorized videos to generate identity fingerprints. This setup is incompatible with our framework, as removing enrollment disables its core mechanism and renders comparison unfair.

5.2 Experimental Results

In this section, we evaluate our method against state-of-the-art baselines for detecting puppeteered videos. We report results under two scenarios: (1) no domain shift—training and testing on all datasets—and (2) cross-domain generalization—training only on NVIDIA VC [11] and evaluating on unseen CREMA-D [55] and RAVDESS [54] datasets.

5.2.1 Performance on Combined Datasets

Tab. 1 summarizes results from training and testing across all datasets (NVIDIA VC [11], CREMA-D [55], and RAVDESS [54]). Our method achieves AUC > 0.97 across all combinations, robustly detecting puppeteering across diverse identities, poses, and expressions. In contrast, deepfake detectors perform poorly; CLRNet [72], the strongest baseline, reaches only 0.684 AUC on MCNET-generated [68] videos from CREMA-D [55]. This highlights the limitation of deepfake detectors, which focus on distinguishing real from synthetic rather than authorized vs. unauthorized identities.

Our method also outperforms the closest prior work, Vahdati et al. [10], reducing relative error by 46% (AUC from 0.945 to 0.971). Notably, our minimum AUC remains above 0.95, compared to 0.90 for Vahdati et al. [10]. This gain stems from our enhanced biometric leakage space, which better separates identity from pose and expression cues. In contrast, their method relies on simple Euclidean distance, making it less robust to such variations.

5.2.2 Cross-Domain Generalization Performance

Tab. 2 presents results when training is limited to NVIDIA VC [11] and testing is performed on CREMA-D [55] and RAVDESS [54]. Our method maintains strong performance with an average AUC of 0.925—only a 5% drop from the no-domain-shift setting—despite training on just 46 identities. This drop is consistent with Fig. 4, which shows AUC scaling with the number of identities used in training. These results suggest the observed performance gap is due to identity diversity rather than differences in appearance or expression, confirming that our method scales with data and generalizes well to unseen domains.

6 Ablation Study

In this section, we conduct a series of ablation experiments to understand the impact of different design choices on the detection performance of our method. To do this, we measured the average detection AUC and relative error reduction (RER) over the self-reenacted and puppeteered examples across all datasets using 3DFaceShop [67] as the generator. The results are provided in Tab. 4.

Biometric Leakage Space Network Design. The results in Tab. 4 show that removing the biometric leakage embedding modules (h_1, h_2) resulted in a substantial drop in performance. This finding confirms that simply comparing z_t to $f(R)$ does not work. Additionally, we observe that increasing the MLP's output dimension to match or exceed the input drops performance to 0.827 AUC (an 80.35% increase in error), indicating that projecting embeddings into a lower-dimensional space effectively filters out irrelevant variability.

Training Protocol. Table 4 highlights three key sensitivities. Limiting each batch to a single positive–negative pair slashes AUC from 0.966 to 0.749 (+86% RER), proving multiple samples are vital for capturing fine biometric cues. Replacing our pose-conditioned contrastive loss with a simple similarity regression lowers AUC to 0.788 (+84% RER), confirming the need to optimise relative identity differences. Finally, retaining extreme-pose frames drops AUC to 0.929 (+52% RER), so filtering them is essential for peak accuracy.

Pose Sensitivity Analysis. We ablated pose sensitivity by progressively omitting frames whose head rotation exceeded a set angle and tracking AUC on a test set. Table 5 shows performance rising steadily as extreme poses are removed, peaking at $\pm 18°$; stricter cut-offs then hurt accuracy. Thus, judicious pose exclusion yields measurable detection gains.

Table 5: Puppeteering detection AUC at increasing yaw thresholds. Performance peaks at $\pm 18°$.

Yaw (°)	± 3	± 8	± 13	± 18
AUC	0.932	0.938	0.953	**0.966**
Yaw (°)	± 23	± 28	± 33	± 38
AUC	0.958	0.942	0.930	0.929

Table 6: Robustness to facial appearance changes on NVIDIA VC (3DFaceShop).

Modification	**AUC**
None (baseline)	0.966
Eyeglasses	0.9393
Piercings	0.9605
Makeup	0.9574

7 Discussion

Computational Efficiency. We benchmarked our method on an RTX 3090 GPU, where it achieved on average 75 FPS—well above the 60 FPS real-time threshold while having under 1M total number of parameters. In comparison, Vahdati et al. [10] reached only 32 FPS under the same conditions. This speed advantage highlights the efficiency of our latent-space analysis, supporting real-time deployment in AI-based videoconferencing.

Temporal Window Size. Fig. 4 shows how detection performance varies with the LSTM's temporal window size. AUC improves sharply from 0.77 at 10 frames to 0.97 at 40 frames, after which gains plateau. A 40-frame window (≈1.3s at 30fps) captures sufficient temporal biometric information, highlighting the value of modeling frame-to-frame dependencies in latent space.

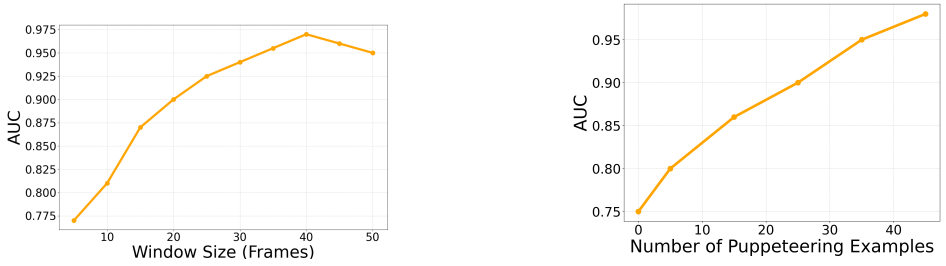

Figure 4: Detection AUC vs. window size and number of puppeteered identities during training.

Scalability. Fig. 4 shows detection AUC as a function of the number of cross-reenacted identities used in training on the NVIDIA VC dataset. Performance steadily increases from 0.75 with few identities to 0.96 with all 46, with no signs of saturation. This suggests that greater diversity during training

enhances the robustness and generalizability of our biometric embedding space in distinguishing puppeteering from self-reenactment.

Robustness to facial appearance changes. We tested our method's robustness to appearance changes by digitally adding eyeglasses, piercings, or heavy makeup to NVIDIA-VC identities using 3DFaceShop. As Table 6 shows, AUC shifted modestly from 0.966 to 0.939 (glasses) and 0.961 (piercings), and even improved to 0.974 with makeup. NVFAIR's natural accessory and cosmetic diversity had already exposed the model to such variations, explaining the small changes and confirming real-world robustness.

Limitations. Our method has three main failure modes. First, extreme head rotation or occlusion (e.g., from hands or objects) can obscure facial features critical to identity. Second, poor lighting or overexposure weakens the biometric signal captured by the camera. Third, motion blur—caused by rapid movement or blinking—can distort the latent representation. Each of these degrades biometric consistency and impairs detection. Examples are provided in the supplementary material.

8 Conclusion

We present a real-time puppeteering defense that authenticates speakers directly in the transmitted pose–expression stream. A low dimensional biometric-leakage embedding, learned with novel pose-conditioned contrastive loss and reinforced by an LSTM for temporal fusion, plus extreme-pose filtering, lifts AUC beyond prior work and stays strong across domains and appearance edits. Ablations confirm every module's value, establishing a practical, robust safeguard for next-generation talking-head videoconferencing systems.

Acknowledgments

This material is based upon work supported by the National Science Foundation under Grant No. 2320600. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.This research was supported in part by NVIDIA Corporation. We thank the NVIDIA Research team for their collaboration, resources, and valuable feedback throughout this project.

References

[1] Hang Chu, Shugao Ma, Fernando De la Torre, Sanja Fidler, and Yaser Sheikh. Expressive telepresence via modular codec avatars. In *European Conference on Computer Vision*, pages 330–345. Springer, 2020.

[2] Guohao Li, Hongyu Yang, Yifang Men, Di Huang, Weixin Li, Ruijie Yang, and Yunhong Wang. Generating Editable Head Avatars with 3D Gaussian GANs. In *2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[3] Matthijs Van Der Boon, Leonor Fermoselle, Frank Ter Haar, Sylvie Dijkstra-Soudarissanane, and Omar Niamut. Deep learning augmented realistic avatars for social VR human representation. In *Proceedings of the 2022 ACM International Conference on Interactive Media Experiences*, pages 311–318, 2022.

[4] Michael Stengel, Koki Nagano, Chao Liu, Matthew Chan, Alex Trevithick, Shalini De Mello, Jonghyun Kim, and David Luebke. AI-Mediated 3D Video Conferencing. In *ACM SIGGRAPH Emerging Technologies*, 2023. doi: 10.1145/3588037.3595385.

[5] Shwetha Rajaram, Nels Numan, Balasaravanan Thoravi Kumaravel, Nicolai Marquardt, and Andrew D. Wilson. BlendScape: Enabling Unified and Personalized Video-Conferencing Environments through Generative AI. *arXiv preprint arXiv:2403.13947*, 2024.

[6] Zhengang Li, Sheng Lin, Shan Liu, Songnan Li, Xue Lin, Wei Wang, and Wei Jiang. FAIVConf: Face Enhancement for AI-based Video Conference with Low Bit-rate. *arXiv preprint arXiv:2207.04090*, 2022.

[7] Ross Cutler, Ramin Mehran, Sam Johnson, Cha Zhang, Adam Kirk, Oliver Whyte, and Adarsh Kowdle. Multimodal Active Speaker Detection and Virtual Cinematography for Video Conferencing. *arXiv preprint arXiv:2002.03977*, 2020.

[8] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. One-Shot Free-View Neural Talking-Head Synthesis for Video Conferencing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10039–10049, 2021. doi: 10.1109/CVPR46437.2021.00991.

[9] Haonan Tong, Haopeng Li, Hongyang Du, Zhaohui Yang, Changchuan Yin, and Dusit Niyato. Multimodal Semantic Communication for Generative Audio-Driven Video Conferencing. *arXiv preprint arXiv:2410.22112*, 2024.

[10] Danial Samadi Vahdati, Tai Duc Nguyen, and Matthew C. Stamm. Defending Low-Bandwidth Talking Head Videoconferencing Systems From Real-Time Puppeteering Attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 983–992, June 2023.

[11] Ekta Prashnani, Koki Nagano, Shalini De Mello, David Luebke, and Orazio Gallo. Avatar fingerprinting for authorized use of synthetic talking-head videos. In *European Conference on Computer Vision*, pages 209–228. Springer, 2024.

[12] Davide Alessandro Coccomini, Nicola Messina, Claudio Gennaro, and Fabrizio Falchi. Combining efficientnet and vision transformers for video deepfake detection. In *International conference on image analysis and processing*, pages 219–229. Springer, 2022.

[13] Nicolo Bonettini, Edoardo Daniele Cannas, Sara Mandelli, Luca Bondi, Paolo Bestagini, and Stefano Tubaro. Video face manipulation detection through ensemble of cnns. In *2020 25th international conference on pattern recognition (ICPR)*, pages 5012–5019. IEEE, 2021.

[14] Yuting Xu, Jian Liang, Gengyun Jia, Ziming Yang, Yanhao Zhang, and Ran He. Tall: Thumbnail layout for deepfake video detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 22658–22668, 2023.

[15] Ying Xu, Kiran Raja, and Marius Pedersen. Supervised contrastive learning for generalizable and explainable deepfakes detection. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 379–389, 2022.

[16] Dat Nguyen, Nesryne Mejri, Inder Pal Singh, Polina Kuleshova, Marcella Astrid, Anis Kacem, Enjie Ghorbel, and Djamila Aouada. Laa-net: Localized artifact attention network for quality-agnostic and generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17395–17405, 2024.

[17] Tai D Nguyen, Shengbang Fang, and Matthew C Stamm. Videofact: detecting video forgeries using attention, scene context, and forensic traces. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 8563–8573, 2024.

[18] Danial Samadi Vahdati, Tai D. Nguyen, Aref Azizpour, and Matthew C. Stamm. Beyond Deepfake Images: Detecting AI-Generated Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4397–4408, June 2024.

[19] Lichuan Ji, Yingqi Lin, Zhenhua Huang, Yan Han, Xiaogang Xu, Jiafei Wu, Chong Wang, and Zhe Liu. Distinguish Any Fake Videos: Unleashing the Power of Large-scale Data and Motion Features. *arXiv preprint arXiv:2405.15343*, 2024.

[20] Chirui Chang, Zhengzhe Liu, Xiaoyang Lyu, and Xiaojuan Qi. What Matters in Detecting AI-Generated Videos like Sora? *arXiv preprint arXiv:2406.19568*, 2024.

[21] Rohit Kundu, Hao Xiong, Vishal Mohanty, Athula Balachandran, and Amit K Roy-Chowdhury. Towards a universal synthetic video detector: From face or background manipulations to fully ai-generated content. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28050–28060, 2025.

[22] Xuangeng Chu and Tatsuya Harada. Generalizable and animatable gaussian head avatar. *Advances in Neural Information Processing Systems*, 37:57642–57670, 2024.

[23] Radek Daněček, Michael J Black, and Timo Bolkart. Emoca: Emotion driven monocular face capture and animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20311–20322, 2022.

[24] Suzhen Wang, Lincheng Li, Yu Ding, Changjie Fan, and Xin Yu. Audio2Head: Audio-driven One-shot Talking-head Generation with Natural Head Motion. In *Proceedings of the 30th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 1107–1113, 2021. doi: 10.24963/ijcai.2021/152.

[25] Xiaoqian Shen, Faizan Farooq Khan, and Mohamed Elhoseiny. Emotalker: Audio driven emotion aware talking head generation. In *Proceedings of the Asian Conference on Computer Vision*, pages 1900–1917, 2024.

[26] Trevine Oorloff and Yaser Yacoob. Expressive Talking Head Video Encoding in StyleGAN2 Latent Space. *arXiv preprint arXiv:2203.14512*, 2022.

[27] Lele Chen, Guofeng Cui, Celong Liu, Zhong Li, Ziyi Kou, Yi Xu, and Chenliang Xu. Talking-Head Generation with Rhythmic Head Motion. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 35–51, 2020.

[28] Madhav Agarwal, Anchit Gupta, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C.V. Jawahar. Compressing Video Calls using Synthetic Talking Heads. In *Proceedings of the British Machine Vision Conference (BMVC)*, pages 1–12, 2021.

[29] Fa-Ting Hong, Longhao Zhang, Li Shen, and Dan Xu. Depth-aware generative adversarial network for talking head video generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3397–3406, 2022.

[30] Fa-Ting Hong, Li Shen, and Dan Xu. Dagan++: Depth-aware generative adversarial network for talking head video generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):2997–3012, 2023.

[31] Arun Mallya, Ting-Chun Wang, and Ming-Yu Liu. Implicit warping for animation with image sets. *Advances in Neural Information Processing Systems*, 35:22438–22450, 2022.

[32] Yue Wu, Yu Deng, Jiaolong Yang, Fangyun Wei, Qifeng Chen, and Xin Tong. Anifacegan: Animatable 3d-aware face image generation for video avatars. *Advances in Neural Information Processing Systems*, 35:36188–36201, 2022.

[33] Hyoung-Kyu Song, Sang Hoon Woo, Junhyeok Lee, Seungmin Yang, Hyunjae Cho, Youseong Lee, Dongho Choi, and Kang-wook Kim. Talking face generation with multilingual tts. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 21425–21430, 2022.

[34] Yue Gao, Yuan Zhou, Jinglu Wang, Xiao Li, Xiang Ming, and Yan Lu. High-fidelity and freely controllable talking head video generation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5609–5619, 2023.

[35] Geumbyeol Hwang, Sunwon Hong, Seunghyun Lee, Sungwoo Park, and Gyeongsu Chae. DisCoHead: Audio-and-Video-Driven Talking Head Generation by Disentangled Control of Head Pose and Facial Expressions. *arXiv preprint arXiv:2303.07697*, 2023.

[36] Haotian Wang, Yuzhe Weng, Yueyan Li, Zilu Guo, Jun Du, Shutong Niu, Jiefeng Ma, Shan He, Xiaoyan Wu, Qiming Hu, Bing Yin, Cong Liu, and Qingfeng Liu. EmotiveTalk: Expressive Talking Head Generation through Audio Information Decoupling and Emotional Video Diffusion. *arXiv preprint arXiv:2411.16726*, 2024.

[37] Mohan Zhou, Yalong Bai, Wei Zhang, Ting Yao, and Tiejun Zhao. Interactive conversational head generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[38] Zhiling Ye, LiangGuo Zhang, Dingheng Zeng, Quan Lu, and Ning Jiang. R2-Talker: Realistic Real-Time Talking Head Synthesis with Hash Grid Landmarks Encoding and Progressive Multilayer Conditioning. *arXiv preprint arXiv:2312.05572*, 2023.

[39] Maxime Oquab, Daniel Haziza, Ludovic Schwartz, Tao Xu, Katayoun Zand, Rui Wang, Peirong Liu, and Camille Couprie. Efficient Conditioned Face Animation Using Frontally-Viewed Embedding. *arXiv preprint arXiv:2203.08765*, 2022.

[40] Pulkit Tandon, Shubham Chandak, Pat Pataranutaporn, Yimeng Liu, Anesu M Mapuranga, Pattie Maes, Tsachy Weissman, and Misha Sra. Txt2Vid: Ultra-low bitrate compression of talking-head videos via text. *IEEE Journal on Selected Areas in Communications*, 41(1):107–118, 2022.

[41] Jason Lawrence, Dan B. Goldman, Supreeth Achar, Gregory Major Blascovich, Joseph G. Desloge, Tommy Fortes, Eric M. Gomez, Sascha Häberling, Hugues Hoppe, Andy Huibers, Claude Knaus, Brian Kuschak, Ricardo Martin-Brualla, Harris Nover, Andrew Ian Russell, Steven M. Seitz, and Kevin Tong. Project Starline: A High-Fidelity Telepresence System. *ACM Transactions on Graphics (Proc. SIGGRAPH Asia)*, 40(6):242:1–242:16, 2021. doi: 10.1145/3478513.3480490.

[42] Gil Knafo and Ohad Fried. FakeOut: Leveraging Out-of-domain Self-supervision for Multi-modal Video Deepfake Detection. *arXiv preprint arXiv:2212.00773*, 2022.

[43] Shu Hu, Yuezun Li, and Siwei Lyu. Exposing GAN-generated faces using inconsistent corneal specular highlights. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2500–2504. IEEE, 2021.

[44] Shruti Agarwal, Hany Farid, Ohad Fried, and Maneesh Agrawala. Detecting deep-fake videos from phoneme-viseme mismatches. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 660–661, 2020.

[45] Hua Qi, Qing Guo, Felix Juefei-Xu, Xiaofei Xie, Lei Ma, Wei Feng, Yang Liu, and Jianjun Zhao. Deeprhythm: Exposing deepfakes with attentional visual heartbeat rhythms. In *Proceedings of the 28th ACM international conference on multimedia*, pages 4318–4327, 2020.

[46] Umur Aybars Ciftci, Ilke Demir, and Lijun Yin. Fakecatcher: Detection of synthetic portrait videos using biological signals. *IEEE transactions on pattern analysis and machine intelligence*, 2020.

[47] Wenxuan Wu, Wenbo Zhou, Weiming Zhang, Han Fang, and Nenghai Yu. Capturing the lighting inconsistency for deepfake detection. In *International Conference on Artificial Intelligence and Security*, pages 637–647. Springer, 2022.

[48] Kaiyue Tian, Chen Chen, Yichao Zhou, and Xiyuan Hu. Illumination enlightened spatial-temporal inconsistency for deepfake video detection. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2024.

[49] Zhimao Lai, Jicheng Li, Chuntao Wang, Jianhua Wu, and Donghua Jiang. LIDeepDet: deepfake detection via image decomposition and advanced lighting information analysis. *Electronics*, 13(22):4466, 2024.

[50] Andrea Ciamarra, Roberto Caldelli, Federico Becattini, Lorenzo Seidenari, and Alberto Del Bimbo. Deepfake detection by exploiting surface anomalies: the SurFake approach. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1024–1033, 2024.

[51] Boyuan Liu, Xin Zhang, Hefei Ling, Zongyi Li, Runsheng Wang, Hanyuan Zhang, and Ping Li. Aim-bone: Texture discrepancy generation and localization for generalized deepfake detection. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 7(3):422–431, 2025.

[52] Zhiqing Guo, Gaobo Yang, Jiyou Chen, and Xingming Sun. Exposing deepfake face forgeries with guided residuals. *IEEE Transactions on Multimedia*, 25:8458–8470, 2023.

[53] Xin Li, Rongrong Ni, Pengpeng Yang, Zhiqiang Fu, and Yao Zhao. Artifacts-disentangled adversarial learning for deepfake detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(4): 1658–1670, 2022.

[54] Steven R Livingstone and Frank A Russo. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PloS one*, 13(5):e0196391, 2018.

[55] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014.

[56] Shuling Zhao, Fa-Ting Hong, Xiaoshui Huang, and Dan Xu. Synergizing Motion and Appearance: Multi-Scale Compensatory Codebooks for Talking Head Video Generation. *arXiv preprint arXiv:2412.00719*, 2024.

[57] Shuai Tan, Bin Ji, Yu Ding, and Ye Pan. Say anything with any style. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5088–5096, 2024.

[58] Haoyu Ma, Tong Zhang, Shanlin Sun, Xiangyi Yan, Kun Han, and Xiaohui Xie. CVTHead: One-shot controllable head avatar with vertex-feature transformer. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 6131–6141, 2024.

[59] Shuai Shen, Wanhua Li, Zheng Zhu, Yueqi Duan, Jie Zhou, and Jiwen Lu. Learning dynamic facial radiance fields for few-shot talking head synthesis. In *European conference on computer vision*, pages 666–682. Springer, 2022.

[60] Wenxuan Zhang, Xiaodong Cun, Xuan Wang, Yong Zhang, Xi Shen, Yu Guo, Ying Shan, and Fei Wang. Sadtalker: Learning realistic 3d motion coefficients for stylized audio-driven single image talking face animation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8652–8661, 2023.

[61] Yaohui Wang, Di Yang, Francois Bremond, and Antitza Dantcheva. Latent image animator: Learning to animate images via latent space navigation. *arXiv preprint arXiv:2203.09043*, 2022.

[62] Madhav Agarwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. Audio-visual face reenactment. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 5178–5187, 2023.

[63] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5265–5274, 2018.

[64] Minchul Kim, Anil K Jain, and Xiaoming Liu. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18750–18759, 2022.

[65] Yunwen Lei, Tianbao Yang, Yiming Ying, and Ding-Xuan Zhou. Generalization analysis for contrastive representation learning. In *International Conference on Machine Learning*, pages 19200–19227. PMLR, 2023.

[66] Daniel Rho, TaeSoo Kim, Sooill Park, Jaehyun Park, and JaeHan Park. Understanding Contrastive Learning Through the Lens of Margins. *arXiv preprint arXiv:2306.11526*, 2023.

[67] Junshu Tang, Bo Zhang, Binxin Yang, Ting Zhang, Dong Chen, Lizhuang Ma, and Fang Wen. 3DFaceShop: Explicitly Controllable 3D-Aware Portrait Generation. *IEEE Transactions on Visualization and Computer Graphics*, 30(9):6020–6037, 2024. doi: 10.1109/TVCG.2023.3323578.

[68] Fa-Ting Hong and Dan Xu. Implicit identity representation conditioned memory compensation network for talking head video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23062–23072, 2023.

[69] Nikita Drobyshev, Antoni Bigata Casademunt, Konstantinos Vougioukas, Zoe Landgraf, Stavros Petridis, and Maja Pantic. Emoportraits: Emotion-enhanced multimodal one-shot head avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8498–8507, 2024.

[70] Stella Bounareli, Vasileios Argyriou, and Georgios Tzimiropoulos. Finding Directions in GAN's Latent Space for Neural Face Reenactment. *British Machine Vision Conference (BMVC)*, 2022.

[71] Jianzhu Guo, Dingyun Zhang, Xiaoqiang Liu, Zhizhou Zhong, Yuan Zhang, Pengfei Wan, and Di Zhang. Liveportrait: Efficient portrait animation with stitching and retargeting control. *arXiv preprint arXiv:2407.03168*, 2024.

[72] Shahroz Tariq, Sangyup Lee, and Simon S Woo. A Convolutional LSTM based Residual Network for Deepfake Video Detection. *arXiv preprint arXiv:2009.07480*, 2020.

NeurIPS Paper Checklist

1. **Claims**
   Answer: [Yes]
   Justification: All scientific claims are stated in Secs. 1–4; empirical claims are validated in Sec. 5 and Tab. 1–4.
2. **Limitations**
   Answer: [Yes]
   Justification: Sec. 7 "Limitations" details failure modes (extreme pose, lighting, motion blur).
3. **Theory assumptions and proofs**
   Answer: [NA]
   Justification: Work is empirical; no formal theorems are presented.
4. **Experimental result reproducibility**
   Answer: [Yes]
   Justification: Hyper-parameters, random seed, and training protocol given in Sec. 5.1; code will be released.
5. **Open access to data and code**
   Answer: [Yes]
   Justification: Public datasets (NVFAIR) are cited; inference/training code and weights will be provided upon acceptance.
6. **Experimental setting/details**
   Answer: [Yes]
   Justification: Dataset splits, generators, and evaluation metrics described in Sec. 5.2–5.3.
7. **Experiment statistical significance**
   Answer: [Yes]
   Justification:
8. **Experiments compute resources**
   Answer: [Yes]
   Justification: Sec. 5.1 states training cost (≈18 GPU-hours on a single RTX 3090) and runtime (75 FPS).
9. **Code of ethics**
   Answer: [Yes]
   Justification: Research follows the NeurIPS Code of Ethics; see Broader-Impact section.
10. **Broader impacts**
    Answer: [Yes]
    Justification: Societal impacts, misuse potential, and mitigation discussed in Broader-Impact paragraph.
11. **Safeguards**
    Answer: [Yes]
    Justification: Release plan includes watermarking detector outputs and open-sourcing under non-commercial license.
12. **Licenses for existing assets**
    Answer: [Yes]
    Justification: We have properly attributed the work of others and have followed licensing and usage terms.
13. **New assets**
    Answer: [NA]
    Justification: No new dataset or proprietary model weights created; only derived metrics.
14. **Crowdsourcing and research with human subjects**
    Answer: [NA]
    Justification: Work uses publicly available datasets; no new human-subject data collected.

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether IRB approvals were obtained?

Answer: [NA]
Justification: Not applicable (no new human-subjects research).

15. **Declaration of LLM usage**
Answer: [NA]
Justification: No large language model is an original or non-standard part of the core methodology.