LEARNING PRIVACY-PRESERVING REPRESENTATIONS FOR AUDIO-VISUAL DEEPFAKE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce Privacy-Preserving Audio-Visual Deepfake Detection (PPDD), which aims to suppress private information while preserving robust and generalizable deepfake detection. Existing privacy-preserving approaches primarily target unimodal tasks and are inadequate for audio-visual deepfake detection, where subtle spatial and temporal manipulation cues may coexist with privacy information, requiring fine-grained privacy suppression. To address this, we propose an end-to-end adversarial framework that learns privacy-preserving audio and visual representations using modality-specific anonymizers. A Feature Suppression Block (FSB) performs fine-grained feature-level suppression, guided by unimodal and cross-modal privacy supervision. Experiments on FakeAVCeleb demonstrate substantial privacy suppression with negligible loss in detection utility. The learned representations further preserve detection utility across unseen manipulations and datasets, transfer across detectors and modalities, suppress unseen privacy attributes, and remain resistant to inversion attacks.

1 INTRODUCTION

Recent advances in generative modeling have enabled highly realistic audio, visual, and multimodal deepfakes, increasing risks of fraud, identity theft, and misinformation. These threats have been addressed by developing unimodal (Haliassos et al., 2021; 2022; Zheng et al., 2021) and multimodal (Oorloff et al., 2024; Anshul et al., 2025a) detectors that generalize to unseen manipulations. They rely on rich facial and speech cues that also encode sensitive information about the speaker. Access to the input or its representations may therefore reveal attributes such as gender, ethnicity, age, and affective states. This risk is especially relevant for third-party or cloud-based services, such as remote identity verification, video-conferencing and content moderation, where audio-visual data must be analyzed to detect impersonation, synthetic identities, or manipulated content. We therefore ask: *can we learn audio-visual representations that suppress private information while preserving the manipulation cues needed for reliable and generalizable deepfake detection?*

To the best of our knowledge, this paper is the first to introduce Privacy-Preserving Audio-Visual Deepfake Detection (PPDD), which aims to detect deepfakes while preventing private-information leakage. The closest related task is Privacy-Preserving Action Recognition (PPAR) (Wu et al., 2020), where private information is suppressed while preserving task-relevant cues. Recent adversarial PPAR methods either operate at frame-level (Dave et al., 2022) leaving temporal privacy cues, or discard video tokens (Li et al., 2023b), which is less suitable for deepfake detection where privacy and subtle manipulation cues may coexist within the same region. More recently, SPLAVU (Fioresi et al., 2026) performs latent-space anonymization within a particular encoder, requiring downstream models to consume its latent features. This can limit compatibility with audio-visual deepfake detectors (Anshul et al., 2025a; 2026; Oorloff et al., 2024) that use their own end-to-end trained encoders. Instead, we seek privacy-preserving outputs in the form of modality-native audio and video that can be used by different detectors without architecture-specific adaptation.

Importantly, we view privacy-preserving deepfake detection as a generalizable representation-learning problem rather than anonymization in a fixed setting. In particular, the learned representations should suppress private information under the following four dimensions of generalization: (i) **cross-manipulation and cross-dataset**, preserving detection utility across unseen manipulations and data distributions; (ii) **detector**, supporting different deepfake detectors than the one used during

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

anonymization training; (iii) **modality**, extending from multimodal to unimodal deepfake detection; and (iv) **privacy-attribute**, suppressing sensitive attributes unseen during anonymization training.

We propose an end-to-end adversarial framework that learns privacy-preserving audio and visual representations while maintaining generalizable deepfake detection. Our framework employs modality-specific anonymizers that produce outputs in the same form as the original audio and video inputs, allowing different deepfake detectors to use the anonymized outputs without architectural adaptation. At the core of each anonymizer, a modality-specific Feature Suppression Block (FSB) operates at the feature level to selectively suppress privacy-related signals, enabling fine-grained control without discarding entire tokens that may also contain critical deepfake artifacts. Guided by unimodal and cross-modal privacy losses, the anonymizers suppress privacy cues within and across modalities, preventing multimodal attackers from exploiting complementary information. Importantly, while multimodal privacy supervision captures cross-modal leakage, the independent audio and video anonymizers also allow their outputs to support unimodal detection. We evaluate privacy leakage across five attributes spanning identity and video-specific affective information: *gender*, *ethnicity*, *age*, *valence*, and *arousal*. Experiments on FakeAVCeleb (Khalid et al., 2021) demonstrate substantial privacy suppression while maintaining strong deepfake detection, with generalization across all four dimensions. In summary:

- We introduce Privacy-Preserving Audio-Visual Deepfake Detection (PPDD) and formulate it as a representation-learning problem that suppresses private information while preserving deepfake detection utility.
- We propose modality-specific anonymizers with a Feature Suppression Block (FSB) that performs fine-grained feature-level suppression of privacy-related information while preserving deepfake cues.
- We jointly optimize unimodal and multimodal privacy losses to suppress both modality-specific and complementary cross-modal privacy cues, ensuring robust privacy protection against multimodal attackers.
- Our method achieves an average reduction of **55%** in privacy leakage while maintaining deepfake detection within **0.6%** AUC of the original model. We also demonstrate generalization of the learned privacy-preserving representations across all four dimensions, together with robustness against black-box and white-box inversion attacks.

2 Related Works

2.1 Audio-Visual Deepfake Detection

Early approaches process audio and visual modalities independently with late fusion (Ilyas et al., 2023; Mittal et al., 2020) or limited cross-modal interaction (Chugh et al., 2020). More recent methods learn joint audio-visual representations through cross-attention (Zhou & Lim, 2021) and contrastive learning (Zou et al., 2024; Yang et al., 2023; Liu et al., 2023). However, single-stage training can overfit to known manipulations, motivating self-supervised approaches for improved generalization (Yu et al., 2023; Cheng et al., 2023). AVAD (Feng et al., 2023) and AVTS (Sung et al., 2023) exploit audio-visual synchronization, while AVFF (Oorloff et al., 2024) uses contrastive masked autoencoding for joint representation learning. PVASS (Yu et al., 2023) models visual-audio alignment, and ICSAVDF (Anshul et al., 2025a) learns intra- and cross-modal temporal synchronization. While effective, these approaches require two-stage training. NFFP (Anshul et al., 2025b) instead performs single-stage intra- and cross-modal feature prediction and achieves strong generalization across manipulations and datasets. Despite these advances, existing detectors operate directly on raw audio-visual data and therefore expose privacy information. We address this complementary problem by learning anonymized audio and visual representations that preserve the cues required for generalizable deepfake detection. Visual-only deepfake detection is discussed in the appendix (Sec A.1).

2.2 Privacy-Preserving Action Recognition

Privacy-preserving action recognition (PPAR) methods broadly use downsampling, obfuscation, or adversarial learning. Downsampling approaches (Chou et al., 2018; Liu & Zhang, 2020; Ryoo et al.,

2017) reduce input resolution to conceal personal information, while simple transformations such as blurring and superpixel clustering have also been explored (Butler et al., 2015). Although these methods require little or no privacy supervision, they often provide a suboptimal privacy-utility trade-off (Dave et al., 2022; Li et al., 2023b). Obfuscation methods (Zhang et al., 2021; Ren et al., 2018) instead detect and modify privacy-sensitive regions, e.g., through face replacement (Ren et al., 2018) or selective blurring (Zhang et al., 2021). However, they require prior knowledge of sensitive regions and can remove information useful for the downstream task.

Adversarial approaches jointly optimize privacy and utility through learning (Wu et al., 2018; 2020) using a min-max strategy to maximize utility while minimizing privacy leakage. Subsequent works explore self-supervised privacy removal (Dave et al., 2022), meta-learning (Peng et al., 2023), and privacy transformations (Kumawat & Nagahara, 2022). Most operate at the frame level, which can preserve exploitable temporal privacy information. STPrivacy (Li et al., 2023b) addresses this using video-level ViT anonymization and token dropping. More recently, SPLAVU (Fioresi et al., 2026) performs privacy suppression in the latent feature space of a pretrained video encoder. However, its anonymized representations are encoder-specific, requiring downstream models to operate on the same latent feature space rather than directly on anonymized video. This limits compatibility with audio-visual deepfake detectors, which typically employ their own encoders.

Existing PPAR methods are video-only. Multimodal privacy-preserving action recognition has considered multiple visual sensors such as RGB, depth, and infrared (Zhang et al., 2023; Rolfsjord et al., 2025), rather than complementary audio-visual information. PPDD instead learns modality-native audio and visual representations with feature suppression, targeting both unimodal and cross-modal privacy leakage while preserving the spatio-temporal cues required for deepfake detection.

2.3 Face Privacy and Privacy-Preserving Deepfake Detection

Facial privacy methods commonly use obfuscation or GAN-based models to conceal identity (Sun et al., 2018; Gafni et al., 2019) or sensitive attributes such as gender, age, and race (Li et al., 2021; Shen et al., 2020; Li et al., 2023a). However, these face-to-face translation methods are not jointly optimized for deepfake detection, which may cause detectors to misclassify real and fake faces. For facial deepfake detection, Chen et al. (2023) employs encryption-based federated learning, but does not evaluate privacy-attribute leakage from the protected representations. In audio, Li et al. (2024; 2026) separate semantic and acoustic information for privacy-preserving deepfake detection, but such disentanglement may retain attributes such as gender and ethnicity. In contrast, PPDD jointly learns audio and visual anonymizers with the deepfake detector to suppress unimodal and cross-modal privacy cues while preserving detection utility.

3 Methodology

3.1 Problem Formulation

Let $\mathcal{D} = \{(v, a, y, \mathbf{s})\}$ denote an audio-visual deepfake dataset, where $v \in \mathcal{V}$ and $a \in \mathcal{A}$ are the video and audio inputs, y is the deepfake label, and $\mathbf{s}$ denotes the privacy attributes. We learn modality-specific anonymizers A_v and A_a that produce privacy-preserving representations

$$\tilde{v} = A_v(v; \theta_v), \qquad \tilde{a} = A_a(a; \theta_a), \tag{1}$$

where $\tilde{v} \in \mathcal{V}$ and $\tilde{a} \in \mathcal{A}$ retain the original input structure and can be directly used by different deepfake detectors.

We define two requirements for these representations. The **utility criterion** requires the deepfake detection loss, L_{DFD}, on anonymized inputs to remain comparable to that on raw inputs:

$$L_{\mathrm{DFD}}\big(D(\tilde{v}, \tilde{a}), y\big) \approx L_{\mathrm{DFD}}\big(D(v, a), y\big), \tag{2}$$

where, D is the deepfake detector. The **privacy criterion** requires private attributes to be harder to infer after anonymization. A single end-to-end privacy attacker P produces audio, video, and cross-modal predictions, capturing both unimodal and complementary audio-visual privacy cues. We therefore require

$$L_{\mathrm{Privacy}}\big(P(\tilde{v}, \tilde{a}), \mathbf{s}\big) \gg L_{\mathrm{Privacy}}\big(P(v, a), \mathbf{s}\big), \tag{3}$$

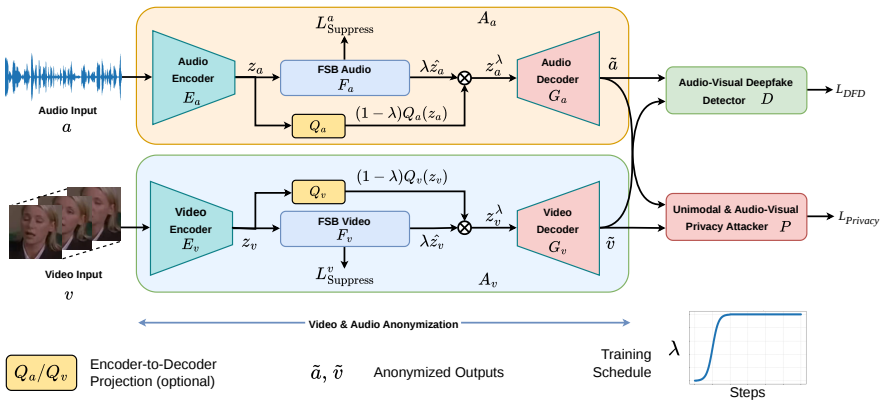


Figure 1: **Overview of PPDD.** Modality-specific anonymizers perform fine-grained feature suppression while preserving deepfake detection utility and reducing unimodal and cross-modal privacy leakage.

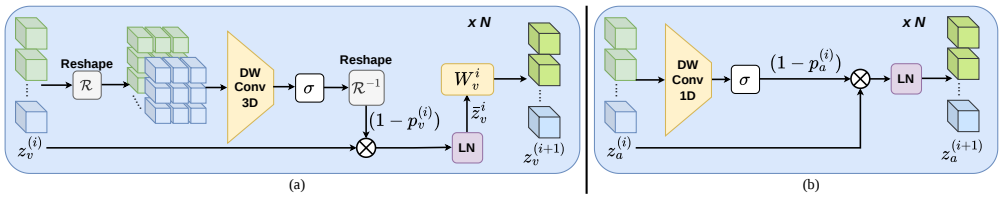


Figure 2: **Feature Suppression Block (FSB).** (a) Video and (b) audio FSBs progressively suppress privacy-related features within each token.

where L_{Privacy} is the privacy supervision loss. Together, these criteria define PPDD as preserving detection utility while suppressing privacy information.

3.2 Modality-Specific Anonymization

The anonymizers are the core components for learning privacy-preserving audio and visual representations. As shown in Fig. 1, the video and audio anonymizers, A_v and A_a, learn privacy-preserving representations under deepfake detection and privacy supervision, and follow an encoder-FSB-decoder architecture. For modality $m \in \{v, a\}$,

$$z_m = E_m(m), \qquad \hat{z}_m = F_m(z_m), \qquad \tilde{m} = G_m(\hat{z}_m), \tag{4}$$

where E_m, F_m, and G_m denote the modality-specific encoder, Feature Suppression Block (FSB), and decoder, respectively. The FSB suppresses privacy information in the latent space, while the decoder produces modality-native outputs $\tilde{v}$ and $\tilde{a}$ that retain the original input structure.

Video and Audio Encoding The raw video v and audio a are first passed through their respective encoders, E_a and E_v, to obtain $z_v \in \mathbb{R}^{n_v \times d_v}$ and $z_a \in \mathbb{R}^{n_a \times d_a}$, where n_m and d_m denote the number and feature dimension of tokens for modality $m \in \{v, a\}$. The resulting z_v and z_a are input to their respective modality-specific FSBs for fine-grained feature suppression.

3.2.1 Feature Suppression Block (FSB)

We introduce FSB, which selectively suppresses privacy-related features in the encoded representation z_m while preserving deepfake-relevant information. Unlike token-level removal (Li et al., 2023b), it learns feature-wise suppression within each token. Each modality-specific FSB refines the supression across N sub-blocks, as shown in Fig. 2. Guided by the competing privacy and de-

tection objectives, it treats the privacy and manipulation cues in the same token separately. The value of N is empirically determined through ablation studies.

Let $z_m^{(i)} \in \mathbb{R}^{n_m \times d_m^{(i)}}$ be the input to the i-th sub-block. We compute its feature-wise suppression map as

$$p_m^{(i)} = \sigma\left(\mathrm{DWConv}_m\left(z_m^{(i)}\right)\right), \qquad p_m^{(i)} \in [0,1]^{n_m \times d_m^{(i)}}, \tag{5}$$

where DWConv_m is a modality-specific depthwise convolution and σ is the sigmoid function. Larger values of $p_m^{(i)}$ indicate stronger suppression. The features are then suppressed as

$$\bar{z}_m^{(i)} = \mathrm{LN}\left[z_m^{(i)} \odot \left(1 - p_m^{(i)}\right)\right], \tag{6}$$

where $\odot$ denotes element-wise multiplication and LN is LayerNorm.

For the **video FSB** (Fig. 2(a)), the 2D token-level representation $z_v^{(i)}$ is reshaped to its 4D spatio-temporal structure before applying a depthwise 3D convolution. This allows the suppression map to capture spatial and temporal context. The suppression map is then reshaped back to the token representation. For the **audio FSB** (Fig. 2(b)), a depthwise 1D convolution operates directly along the temporal dimension.

Each video sub-block also progressively projects the suppressed features toward the decoder dimension as $z_v^{(i+1)} = W_v^{(i)} \bar{z}_v^{(i)}$, using the geometric ratio $r = (d_v^{\mathrm{dec}}/d_v)^{1/N}$. Here, d_v^{dec} denotes the input feature dimension of the video decoder. This avoids abrupt dimensionality reduction while refining suppression across sub-blocks. No projection is required for audio since its encoder and decoder dimensions are equal. The final FSB output gives $\hat{z}_m = F_m(z_m)$, which is passed to the decoder.

Video and Audio Decoding The suppressed representations $\hat{z}_v$ and $\hat{z}_a$ are passed through their respective decoders, G_v and G_a, to obtain the privacy-preserving video $\tilde{v} \in \mathbb{R}^{C \times T_v \times H \times W}$ and audio $\tilde{a} \in \mathbb{R}^{T_a}$, that can be directly used by different unimodal and multimodal deepfake detectors.

3.3 Privacy-Utility Objectives and Controlled Suppression

The anonymizers are guided by two competing objectives: preserving deepfake detection utility and suppressing privacy information.

Deepfake Detection. Given the audio-visual deepfake detector D, its loss on the anonymized representations is $L_{\mathrm{DFD}} = L_{\mathrm{det}}\big(D(\tilde{v}, \tilde{a}), y\big)$, where L_{det} denotes the detector-specific loss.

Privacy Suppression. The privacy attacker P predicts attributes from audio, video, and cross-modal representation. For each modality $m \in \{a, v, av\}$, the privacy loss is

$$L_{\mathrm{Privacy}}^m = \frac{1}{K_m} \sum_{k=1}^{K_m} L_{\mathrm{CE}}\big(\hat{s}_k^m, s_k^m\big), \tag{7}$$

where K_m is the number of predicted privacy attributes, and $\hat{s}_k^m$ and s_k^m are the predicted and ground-truth values of the k-th attribute. The unimodal outputs capture modality-specific privacy leakage, while the audio-visual output captures complementary cross-modal privacy cues. The overall privacy loss is $L_{\mathrm{Privacy}} = \frac{\Sigma w_i L_{\mathrm{Privacy}}^i}{\Sigma w_i}$, where $i \in \{a, v, av\}$, and w_is control the contribution of each output.

Controlled Feature Suppression. The privacy objective determines which features to suppress but does not control the amount of suppression. We therefore regularize the suppression maps using the geometric ratio r, shared by both modalities:

$$L_{\mathrm{Suppress}}^m = \frac{1}{N} \sum_{i=1}^{N} \left(\mathrm{Mean}\left(p_m^{(i)}\right) - r\right)^2, \qquad m \in \{v, a\}. \tag{8}$$

This prevents over- or under-suppression while allowing the privacy and detection objectives to determine which features are suppressed.

The audio encoder additionally uses a discrete codebook, and has its associated commitment loss L_{commit}. The overall anonymizer loss is

$$L_{\text{Anony}} = \frac{L^v_{\text{Suppress}} + L^a_{\text{Suppress}}}{2} + \alpha L_{\text{commit}}, \tag{9}$$

where $\alpha = 0.1$ keeps the commitment and suppression losses on comparable scales.

3.4 Adversarial Privacy-Utility Learning

Initialization. The anonymizer encoders and decoders are initialized from reconstruction-pretrained models, providing an information-preserving starting point from which privacy cues can be selectively suppressed. The FSBs are initialized for minimal suppression, while the deepfake detector and privacy attacker are first trained on raw audio-visual inputs.

Adversarial Training. We alternate between two optimization steps. First, the anonymizers are updated while the detector and privacy attacker are frozen by minimizing

$$L_A = L_{\text{Anony}} + L_{\text{DFD}} - L_{\text{Privacy}}. \tag{10}$$

Thus, the anonymizers preserve deepfake utility by minimizing L_{DFD} while suppressing privacy information by maximizing L_{Privacy}. Second, the anonymizers are frozen and the detector and privacy attacker are updated by minimizing

$$L_{D,P} = L_{\text{DFD}} + L_{\text{Privacy}}. \tag{11}$$

This alternating optimization improves the detector and privacy attacker on the current anonymized representations while training the anonymizers against an adaptive privacy adversary.

Since the FSBs are learned from scratch while encoders and decoders are reconstruction-pretrained, directly replacing the encoder features with FSB outputs can destabilize early training. We therefore gradually introduce the suppressed representations to the decoders. During this transition, the decoder input is

$$z^\lambda_m(t) = \lambda(t)\hat{z}_m + \big(1 - \lambda(t)\big)Q_m(z_m), \tag{12}$$

where Q_m maps the encoder representation to the decoder input dimension and is identity when no projection is required. The coefficient $\lambda(t)$ follows a smooth $\tanh$ schedule from 0 to 1, gradually transitioning the decoder from the original encoder representation to the FSB output. During this transition, $\tilde{m} = G_m(z^\lambda_m(t))$. Once $\lambda(t) = 1$, $z^\lambda_m(t) = \hat{z}_m$, recovering the final anonymization pipeline $\tilde{m} = G_m(\hat{z}_m)$ defined in Eq. 4.

Evaluation. After adversarial training, the anonymizers are frozen. The deepfake detector and privacy attacker are reinitialized and trained from scratch on the anonymized outputs $(\tilde{v}, \tilde{a})$. Their performance is compared with models trained on raw inputs to measure the deepfake detection utility and remaining privacy information. An end-to-end training and evaluation algorithm is presented in appendix (Alg. 1)

4 Experiments

4.1 Experimental Setup

Datasets and Privacy Attributes. We conduct initialization, adversarial training, and evaluation on FakeAVCeleb (Khalid et al., 2021), augmenting its real samples with VoxCeleb2 (Chung et al., 2018) to balance real and fake videos. Gender labels are provided by the dataset, while ethnicity labels are obtained from VMER (Greco et al., 2020). Age is estimated using DeepFace (Serengil & Ozpinar, 2020) for video and VanPy (Koushnir et al., 2025) for audio. Video valence and arousal are estimated frame-wise using EmoNet (Toisoul et al., 2021) and averaged per video, while their audio counterparts are extracted using (Wagner et al., 2023). Continuous age and affect values are quantized into three balanced categories (low, medium, high); comparisons with alternative age discretizations are provided in the appendix (Sec A.4.9).

We consider all five attributes for the unimodal audio and video privacy outputs. For the multimodal output, we use only gender and ethnicity, which are directly available or consistently annotated across modalities, to avoid modality-specific noise from pretrained attribute estimators.

Table 1: **Intra-dataset privacy-utility evaluation on FakeAVCeleb.** We report deepfake detection AUC (↑) and privacy cMAP (↓) for 'Audio', 'Video', and 'Audio-Visual' outputs. 'Avg.' denotes the mean cMAP across the three privacy outputs. $\downarrow x$ denotes the relative reduction in cMAP from raw inputs. **Best** and second-best results are highlighted among privacy-preserving methods.

Method	AUC↑	Audio↓	Video↓	Audio-Visual↓	Avg.↓
Raw	96.30	92.84	88.19	99.60	93.54
DS(56, 2)	92.49	92.15	87.37	99.76	93.09
DS(28, 4)	92.74	92.51	87.76	99.67	93.31
WeakBlur	89.23	92.66	87.30	99.66	93.21
StrongBlur	90.53	93.15	87.53	58.36	79.68
SPAct (Dave et al., 2022)	89.50	76.43	35.40	34.66	48.83
SPAct-Sup (Dave et al., 2022)	95.12	73.41	35.32	37.66	48.80
STPrivacy (Li et al., 2023b)	94.21	72.46	35.62	36.51	48.20
SPLAVU (Fioresi et al., 2026)	**98.40**	58.9	87.5	96.6	81.0
PPDD w/o FSB (Ours)	94.46	63.59↓ 32%	**34.77**↓ 61%	35.62↓ 64%	44.66↓ 52%
PPDD (Ours)	96.24	**56.95**↓ 39%	35.21↓ 60%	**33.71**↓ 66%	**41.96**↓ 55%

Model Configuration. We use $N = 3$ FSB sub-blocks, selected through ablation (Sec A.4.8). The video FSB uses a depthwise 3D convolution with kernel $(3, 5, 5)$, while the audio FSB uses a depthwise 1D convolution with kernel size 5. Since the video decoder dimension is half the encoder dimension, the geometric suppression ratio is $r = (1/2)^{1/3}$.

We use NFFP (Anshul et al., 2025b) as the primary audio-visual deepfake detector, with $L_{\text{DFD}} = L_{\text{BCE}} + L_{\text{con}}$, where L_{con} is averaged across audio, video, and audio-visual representations. The privacy attacker is adapted from AV-HuBERT (Shi et al., 2022) using a ConvTransformer video encoder, a ViT-Tiny (Dosovitskiy, 2021) audio encoder, and attribute-specific heads.

Training and Evaluation. Initialization and adversarial training use 64-frame clips to balance temporal context and computational cost. Privacy evaluation also uses 64 frames, while deepfake detection uses 256 frames to provide richer temporal context. Further dataset, training, and implementation details are provided in the appendix (Sec A.2, A.3).

4.2 Experimental Results

Following (Fioresi et al., 2026), we use class-wise mean average precision (cMAP) to measure privacy leakage, where lower values indicate stronger privacy preservation. For deepfake detection, we report AUC for intra-dataset evaluation and AP/AUC for cross-manipulation generalization (Anshul et al., 2025b).

Intra-Dataset Privacy-Utility Evaluation. We first evaluate the privacy-utility trade-off on FakeAVCeleb by comparing PPDD with conventional perturbation and privacy-preserving baselines. As shown in Tab. 1, our method reduces the average privacy cMAP from 93.54 on raw inputs to 41.96, while decreasing detection AUC by only 0.06%. In contrast, downsampling and weak blurring provide little privacy protection, while stronger blurring improves privacy at a substantial cost to detection performance.

For comparison with adversarial privacy-preserving methods, we adapt SPAct (Dave et al., 2022), STPrivacy (Li et al., 2023b), and SPLAVU (Fioresi et al., 2026) (see implementation details in Sec. A.2) to the audio-visual deepfake setting. For a fair comparison, all methods use the same deepfake detector and privacy attacker, with ViT-Small (Dosovitskiy, 2021) and WavTokenizer (Ji et al., 2025) as the video and audio anonymizer backbones. SPAct retains its frame-level design for both modalities, and we report its original self-supervised variant (SPAct) and a supervised variant using our privacy attacker (SPAct-Sup). We also include an "PPDD w/o FSB" baseline that uses the same encoder-decoder backbones without FSBs.

Table 2: **Cross-manipulation generalization on FakeAVCeleb.** Models are trained while holding out one manipulation category and evaluated on the unseen category. AP and AUC are reported (↑). The five categories are: (i) **RVFA:** Real Visual – Fake Audio by SV2TTS Jia et al. (2018), (ii) **FVRA-WL:** Fake Visual – Real Audio by Wav2Lip Prajwal et al. (2020), (iii) **FVFA-GAN:** Fake Visual – Fake Audio by FaceSwapGAN Nirkin et al. (2019), Wav2Lip, and SV2TTS, (iv) **FVFA-WL:** Fake Visual – Fake Audio by Wav2Lip and SV2TTS, and (v) **FVFA-FS:** Fake Visual – Fake Audio by FaceSwap Korshunova et al. (2017), Wav2Lip, and SV2TTS. AVG-FV denotes the average over the four fake-visual categories.

Method	RVFA		FVRA-WL		FVFA-GAN		FVFA-WL		FVFA-FS		AVG-FV	
	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC
AV-DFD (Zhou & Lim, 2021)	74.9	73.3	97.0	97.4	58.4	55.4	**100.0**	**100.0**	99.6	99.7	88.8	88.1
AVAD (LRS2) (Feng et al., 2023)	62.4	71.6	93.6	93.7	94.1	94.1	93.8	94.1	95.3	95.8	94.2	94.5
AVAD (LRS3) (Feng et al., 2023)	70.7	80.5	91.1	93.0	91.6	92.7	91.4	93.1	91.0	92.3	91.3	92.8
AVFF (Oorloff et al., 2024)	93.3	92.4	94.8	**98.2**	**99.9**	**100.0**	99.4	99.8	**100.0**	**100.0**	98.5	**99.5**
ICSAVDF (Anshul et al., 2025a)	98.6	97.9	-	-	99.6	96.7	99.7	97.4	99.6	98.1	**99.6**	97.4
NFFP (Anshul et al., 2025b)	96.4	95.9	98.5	88.6	99.7	97.8	99.7	97.6	99.7	98.4	99.4	95.6
Ours (NFFP + PPDD)	**99.8**	**99.8**	**99.4**	95.4	**99.9**	**100.0**	99.7	99.9	99.3	99.6	**99.6**	98.7

SPAct achieves strong privacy suppression but substantially reduces detection performance, while SPAct-Sup improves the privacy-utility trade-off, showing the benefit of supervised unimodal and multimodal privacy guidance. STPrivacy further reduces privacy leakage but lowers detection AUC, likely because token removal can discard manipulation cues together with privacy information. SPLAVU achieves the highest detection AUC due to its strong frozen pretrained encoders. However, it provides very limited privacy suppression. Its self-supervised objective suppresses information shared across frames, which may be less suited to deepfake detection than action recognition, as manipulation cues may also persist across frames. Consequently, its privacy objective may compete with preserving detection-relevant information, resulting in a weaker privacy-utility trade-off. "PPDD w/o FSB" provides a stronger privacy-utility balance, which is further improved by the proposed FSB. Overall, PPDD provides the strongest privacy-utility trade-off among the evaluated methods. Comparisons with additional modality-specific baselines are present in appendix (Sec A.4.1). Moreover, we study the privacy-utility trade-off by varying the weight of L_{Privacy} during anonymizer training in the appendix (Sec. A.4.11).

Cross-Manipulation Generalization. We evaluate whether the learned privacy-preserving representations retain detection utility on unseen manipulation types. Following Anshul et al. (2025a); Feng et al. (2023), we use a leave-one-manipulation-out protocol over the five categories in Tab. 2, keeping the anonymizer frozen. As shown in Tab. 2, NFFP trained on PPDD representations maintains its strong cross-manipulation generalization. Interestingly, the performance is slightly higher overall after anonymization. We believe this may be because adversarial representation learning suppresses information that is less relevant to deepfake detection and may otherwise affect generalization. Overall, these results indicate that the learned privacy-preserving representations retain the cross-manipulation generalization ability of the underlying detector.

Cross-Dataset Generalization. To further assess detector generalization, we evaluate PPDD in a cross-dataset setting following prior work (Anshul et al., 2025a;b), training on FakeAVCeleb and testing on a subset of KoDF (Kwon et al., 2021). As shown in Tab. 3(a), PPDD retains strong performance under dataset shift and outperforms AVFF (Oorloff et al., 2024), demonstrating that the learned privacy-preserving representations preserve substantial cross-dataset detection utility.

Novel Privacy-Attribute Generalization. Since PPDD uses supervised privacy losses, we examine whether privacy suppression generalizes beyond the attributes used for training. We train an *Ours-GE* variant using only gender and ethnicity for privacy supervision and evaluate leakage exclusively on the three unseen attributes: age, valence, and arousal. Since the multimodal privacy head predicts only gender and ethnicity, we report audio and video leakage for these attributes. As shown in Tab. 3(c), Ours-GE performs comparably to the full model, indicating that the learned representation can suppress privacy attributes unseen during training. Additional evaluations of cross-dataset privacy suppression (Sec. A.4.2) and transfer to different (Sec. A.4.4) and unimodal (Sec. A.4.3) detectors are provided in the appendix.

Table 3: **Generalization, component analysis, and robustness.** (a) Cross-dataset detection generalization from FakeAVCeleb to KoDF. (b) Component analysis of FSB and suppression regularization. (c) Novel privacy-attribute generalization. (d) Resistance to inversion attacks.

(a) Cross-Dataset Generalization

Method	AP↑	AUC↑
AV-DFD ICCV'21	79.6	82.1
AVAD CVPR'23	87.6	86.9
AVFF CVPR'24	93.1	95.5
ICSAVDF ICCV'25	98.9	99.0
NFFP Arxiv'25	**100.0**	**100.0**
Ours	94.6	95.8

(b) Component Analysis

Method	AUC↑	Avg. cMAP↓
Enc-Dec	94.46	44.66
+ FSB	95.13	42.94
+ L_{Suppress} (Ours)	**96.24**	**41.96**

(c) Novel Attribute Generalization

Method	cMAP↓		
	A	V	Avg.
Raw	94.3	80.5	87.4
Ours-GE	**61.7**	37.2	49.5
Ours	63.3	**34.1**	**48.7**

(d) Inversion Resistance

Attack	PSNR↓		Avg. cMAP↓
	V	A	
Raw	-	-	93.5
Black-box	12.0	14.3	33.5
White-box	10.9	14.3	37.1

Impact of FSB and Suppression Loss. We ablate the two key components of controlled feature suppression: the FSB and its suppression loss. As shown in Tab. 3(b), adding the FSB improves both privacy preservation and detection utility, with further gains after introducing L_{Suppress}, demonstrating the benefit of explicitly controlling feature suppression. Additional ablation studies analyzing the impact of the multimodal privacy loss (Sec. A.4.5), geometric ratio r (Sec. A.4.6), and λ schedule (Sec. A.4.7) are reported in appendix.

Resistance to Inversion Attacks. Following Fioresi et al. (2026), we evaluate whether original audio and video inputs can be recovered from the learned privacy-preserving representations under black-box and white-box inversion attacks. In the black-box setting, the attacker observes raw inputs and their anonymized representations and trains independent inversion models, using a ViT-Base encoder-decoder for video and a WavTokenizer-based model for audio. In the white-box setting, the attacker has full access to the anonymizer and uses gradient-based optimization to recover inputs whose representations match the observed ones. We measure reconstruction fidelity using PSNR and privacy leakage from reconstructed inputs using the privacy attacker. As shown in Tab. 3(d), both attacks yield low PSNR and substantially lower privacy cMAP than raw inputs, indicating resistance to recovering the original inputs and their private information. Moreover, qualitative analysis in appendix (Sec. A.4.10) shows that the anonymizers' outputs contain no human-interpretable information.

5 Conclusion

We introduced Privacy-Preserving Audio-Visual Deepfake Detection (PPDD) and formulated it as a representation-learning problem that aims to suppress private information while preserving deepfake detection utility. We proposed an end-to-end adversarial framework with modality-specific anonymizers and fine-grained feature suppression, producing privacy-preserving representations in the original audio and video domains. Experiments demonstrate substantial privacy suppression with minimal loss in detection utility. Importantly, the learned representations preserve detection utility across unseen manipulations and datasets, different detector architectures, and unimodal settings, while privacy suppression extends to attributes unseen during training. The representations also remain resistant to black-box and white-box inversion attacks. These results demonstrate that PPDD learns transferable privacy-preserving representations whose privacy and utility extend beyond the training configuration. Future work can explore broader privacy attributes and more diverse audio-visual deepfake datasets.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing, improving clarity and presentation, and modifying figures. They were not used to design the proposed framework, execute the experimental pipeline, or generate experimental results. All technical content, experimental design, analysis, and conclusions were developed and verified by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Ashutosh Anshul, Shreyas Gopal, Deepu Rajan, and Eng Siong Chng. Intra-modal and cross-modal synchronization for audio-visual deepfake detection and temporal localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 13826–13836, 2025a.

Ashutosh Anshul, Shreyas Gopal, Deepu Rajan, and Eng Siong Chng. Next-frame feature prediction for multimodal deepfake detection and temporal localization. *arXiv preprint arXiv:2511.10212*, 2025b.

Ashutosh Anshul, Eng Siong Chng, and Deepu Rajan. Av representation learning via audio shift prediction for multimodal deepfake detection and temporal localization. In *2026 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 2553–2563. IEEE, 2026.

Daniel J Butler, Justin Huang, Franziska Roesner, and Maya Cakmak. The privacy-utility tradeoff for remotely teleoperated robots. In *Proceedings of the tenth annual ACM/IEEE international conference on human-robot interaction*, pp. 27–34, 2015.

Zhixi Cai, Shreya Ghosh, Kalin Stefanov, Abhinav Dhall, Jianfei Cai, Hamid Rezatofighi, Reza Haffari, and Munawar Hayat. Marlin: Masked autoencoder for facial video representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1493–1504, June 2023.

Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pp. 67–74. IEEE, 2018.

Beijing Chen, Xin Liu, Zhihua Xia, and Guoying Zhao. Privacy-preserving deepfake face image detection. *Digital Signal Processing*, 143:104233, 2023.

Zongmei Chen, Xin Liao, Xiaoshuai Wu, and Yanxiang Chen. Compressed deepfake video detection based on 3d spatiotemporal trajectories. In *2024 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 1–8, 2024.

Harry Cheng, Yangyang Guo, Tianyi Wang, Qi Li, Xiaojun Chang, and Liqiang Nie. Voice-face homogeneity tells deepfake. *ACM Transactions on Multimedia Computing, Communications and Applications*, 20(3):1–22, 2023.

Jongwook Choi, Taehoon Kim, Yonghyun Jeong, Seungryul Baek, and Jongwon Choi. Exploiting style latent flows for generalizing deepfake detection video detection. *arXiv preprint arXiv:2403.06592*, 2024.

Edward Chou, Matthew Tan, Cherry Zou, Michelle Guo, Albert Haque, Arnold Milstein, and Li Fei-Fei. Privacy-preserving action recognition for smart hospitals using low-resolution depth images. *arXiv preprint arXiv:1811.09950*, 2018.

Komal Chugh, Parul Gupta, Abhinav Dhall, and Ramanathan Subramanian. Not made for each other-audio-visual dissonance-based deepfake detection and localization. In *Proceedings of the 28th ACM international conference on multimedia*, pp. 439–447, 2020.

J. S. Chung, A. Nagrani, and A. Zisserman. Voxceleb2: Deep speaker recognition. In *INTERSPEECH*, 2018.

Davide Cozzolino, Andreas Rössler, Justus Thies, Matthias Nießner, and Luisa Verdoliva. Id-reveal: Identity-aware deepfake video detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 15108–15117, 2021.

Ishan Rajendrakumar Dave, Chen Chen, and Mubarak Shah. Spact: Self-supervised privacy preservation for action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 20164–20173, 2022.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Chao Feng, Ziyang Chen, and Andrew Owens. Self-supervised video forensics by audio-visual anomaly detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10491–10503, 2023.

Joseph Fioresi, Ishan Rajendrakumar Dave, and Mubarak Shah. Privacy beyond pixels: Latent anonymization for privacy-preserving video understanding. In *International Conference on Learning Representations*, volume 2026, pp. 104423–104447, 2026.

Oran Gafni, Lior Wolf, and Yaniv Taigman. Live face de-identification in video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9378–9387, 2019.

Antonio Greco, Gennaro Percannella, Mario Vento, and Vincenzo Vigilante. Benchmarking deep network architectures for ethnicity recognition using a new large face dataset. *Machine Vision and Applications*, 2020.

Alexandros Haliassos, Konstantinos Vougioukas, Stavros Petridis, and Maja Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5039–5049, 2021.

Alexandros Haliassos, Rodrigo Mira, Stavros Petridis, and Maja Pantic. Leveraging real talking faces via self-supervision for robust forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14950–14962, 2022.

Baojin Huang, Zhongyuan Wang, Jifan Yang, Jiaxin Ai, Qin Zou, Qian Wang, and Dengpan Ye. Implicit identity driven deepfake face swapping detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4490–4499, 2023.

Hafsa Ilyas, Ali Javed, and Khalid Mahmood Malik. Avfakenet: A unified end-to-end dense swin transformer deep learning model for audio–visual deepfakes detection. *Applied Soft Computing*, 136:110124, 2023.

Shengpeng Ji, Ziyue Jiang, Wen Wang, Yifu Chen, Minghui Fang, Jialong Zuo, Qian Yang, Xize Cheng, Zehan Wang, Ruiqi Li, et al. Wavtokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling. In *International Conference on Learning Representations*, 2025.

Ye Jia, Yu Zhang, Ron Weiss, Quan Wang, Jonathan Shen, Fei Ren, Patrick Nguyen, Ruoming Pang, Ignacio Lopez Moreno, Yonghui Wu, et al. Transfer learning from speaker verification to multispeaker text-to-speech synthesis. *Advances in neural information processing systems*, 31, 2018.

Hasam Khalid, Shahroz Tariq, Minha Kim, and Simon S. Woo. FakeAVCeleb: A novel audio-video multimodal deepfake dataset. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

Iryna Korshunova, Wenzhe Shi, Joni Dambre, and Lucas Theis. Fast face-swap using convolutional neural networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 3677–3685, 2017.

Gregory Koushnir, Michael Fire, Galit Fuhrmann Alpert, and Dima Kagan. Vanpy: Voice analysis framework, 2025.

Sudhakar Kumawat and Hajime Nagahara. Privacy-preserving action recognition via motion difference quantization. In *Computer Vision – ECCV 2022*. Springer Nature Switzerland, 2022.

Patrick Kwon, Jaeseong You, Gyuhyeon Nam, Sungwoo Park, and Gyeongsu Chae. Kodf: A large-scale korean deepfake detection dataset. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10744–10753, 2021.

Jingzhi Li, Hua Zhang, Siyuan Liang, Pengwen Dai, and Xiaochun Cao. Privacy-enhancing face obfuscation guided by semantic-aware attribution maps. *IEEE Transactions on Information Forensics and Security*, 18:3632–3646, 2023a.

Lingzhi Li, Jianmin Bao, Ting Zhang, Hao Yang, Dong Chen, Fang Wen, and Baining Guo. Face x-ray for more general face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5001–5010, 2020.

Ming Li, Xiangyu Xu, Hehe Fan, Pan Zhou, Jun Liu, Jia-Wei Liu, Jiahe Li, Jussi Keppo, Mike Zheng Shou, and Shuicheng Yan. Stprivacy: Spatio-temporal privacy-preserving action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5106–5115, 2023b.

Xinfeng Li, Kai Li, Yifan Zheng, Chen Yan, Xiaoyu Ji, and Wenyuan Xu. Safeear: Content privacy-preserving audio deepfake detection. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pp. 3585–3599, 2024.

Xinfeng Li, Yifan Zheng, Chen Yan, Kai Li, Chang Zeng, Xiaoyu Ji, and Wenyuan Xu. Critical information only: A content privacy-preserving framework for detecting audio deepfakes. *IEEE Transactions on Dependable and Secure Computing*, 23(2):2165–2182, 2026. doi: 10.1109/TDSC.2025.3624972.

Xinyang Li, Shengchuan Zhang, Jie Hu, Liujuan Cao, Xiaopeng Hong, Xudong Mao, Feiyue Huang, Yongjian Wu, and Rongrong Ji. Image-to-image translation via hierarchical style disentanglement. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8639–8648, 2021.

Li Lin, Xinan He, Yan Ju, Xin Wang, Feng Ding, and Shu Hu. Preserving fairness generalization in deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16815–16825, 2024.

Decheng Liu, Zeyang Zheng, Chunlei Peng, Yukai Wang, Nannan Wang, and Xinbo Gao. Hierarchical forgery classifier on multi-modality face forgery clues. *IEEE Transactions on Multimedia*, 26:2894–2905, 2024.

Jixin Liu and Leilei Zhang. Indoor privacy-preserving action recognition via partially coupled convolutional neural network. In *2020 International Conference on Artificial Intelligence and Computer Engineering (ICAICE)*, pp. 292–295. IEEE, 2020.

Xiaolong Liu, Yang Yu, Xiaolong Li, and Yao Zhao. Mcl: multimodal contrastive learning for deepfake detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(4): 2803–2813, 2023.

Zhengzhe Liu, Xiaojuan Qi, and Philip HS Torr. Global texture enhancement for fake face detection in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8060–8069, 2020.

Trisha Mittal, Uttaran Bhattacharya, Rohan Chandra, Aniket Bera, and Dinesh Manocha. Emotions don't lie: An audio-visual deepfake detection method using affective cues. In *Proceedings of the 28th ACM international conference on multimedia*, pp. 2823–2832, 2020.

Yuval Nirkin, Yosi Keller, and Tal Hassner. Fsgan: Subject agnostic face swapping and reenactment. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7184–7193, 2019.

Allassan Tchangmena A Nken, Susan McKeever, Peter Corcoran, and Ihsan Ullah. Video-dprp: A differentially private approach for visual privacy-preserving video human activity recognition. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 345–362. Springer, 2025.

Trevine Oorloff, Surya Koppisetti, Nicolò Bonettini, Divyaraj Solanki, Ben Colman, Yaser Yacoob, Ali Shahriyari, and Gaurav Bharaj. Avff: Audio-visual feature fusion for video deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27102–27112, 2024.

Duo Peng, Li Xu, Qiuhong Ke, Ping Hu, and Jun Liu. Joint attribute and model generalization learning for privacy-preserving action recognition. *Advances in Neural Information Processing Systems*, 36:58704–58716, 2023.

KR Prajwal, Rudrabha Mukhopadhyay, Vinay P Namboodiri, and CV Jawahar. A lip sync expert is all you need for speech to lip generation in the wild. In *Proceedings of the 28th ACM international conference on multimedia*, pp. 484–492, 2020.

Ekta Prashnani, Michael Goebel, and B. S. Manjunath. Generalizable deepfake detection with phase-based motion analysis. *IEEE Transactions on Image Processing*, 34:100–112, 2025.

Tong Qiao, Shichuang Xie, Yanli Chen, Florent Retraint, Ran Shi, and Xiangyang Luo. Deepfake detection fighting against noisy label attack. *IEEE Transactions on Multimedia*, 26:9047–9059, 2024.

Zhongzheng Ren, Yong Jae Lee, and Michael S Ryoo. Learning to anonymize faces for privacy preserving action detection. In *Proceedings of the european conference on computer vision (ECCV)*, pp. 620–636, 2018.

Sigmund Rolfsjord, Safia Fatima, Hugh Alexander von Arnim, and Adel Baselizadeh. Multimodal transfer learning for privacy in human activity recognition. In *2025 34th IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, pp. 15–20. IEEE, 2025.

Michael Ryoo, Brandon Rothrock, Charles Fleming, and Hyun Jong Yang. Privacy-preserving human activity recognition from extreme low resolution. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

Sefik Ilkin Serengil and Alper Ozpinar. Lightface: A hybrid deep face recognition framework. In *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*, pp. 23–27. IEEE, 2020.

Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE transactions on pattern analysis and machine intelligence*, 44(4):2004–2018, 2020.

Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. *arXiv preprint arXiv:2201.02184*, 2022.

Qianru Sun, Liqian Ma, Seong Joon Oh, Luc Van Gool, Bernt Schiele, and Mario Fritz. Natural and effective obfuscation by head inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5050–5059, 2018.

Chang-Sung Sung, Jun-Cheng Chen, and Chu-Song Chen. Hearing and seeing abnormality: Self-supervised audio-visual mutual learning for deepfake detection. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023.

Chuangchuang Tan, Yao Zhao, Shikui Wei, Guanghua Gu, Ping Liu, and Yunchao Wei. Rethinking the up-sampling operations in cnn-based generative network for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28130–28139, 2024.

Antoine Toisoul, Jean Kossaifi, Adrian Bulat, Georgios Tzimiropoulos, and Maja Pantic. Estimation of continuous valence and arousal levels from faces in naturalistic conditions. *Nature Machine Intelligence*, 2021.

Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W Schuller. Dawn of the transformer era in speech emotion recognition: closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10745–10759, 2023.

Zhendong Wang, Jianmin Bao, Wengang Zhou, Weilun Wang, and Houqiang Li. Altfreezing for more general video face forgery detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 4129–4138, 2023.

Zhi Wang, Yiwen Guo, and Wangmeng Zuo. Deepfake forensics via an adversarial game. *IEEE Transactions on Image Processing*, 31:3541–3552, 2022.

Jianghao Wu, Baopeng Zhang, Zhaoyang Li, Guilin Pang, Zhu Teng, and Jianping Fan. Interactive two-stream network across modalities for deepfake detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(11):6418–6430, 2023.

Zhenyu Wu, Zhangyang Wang, Zhaowen Wang, and Hailin Jin. Towards privacy-preserving visual recognition via adversarial training: A pilot study. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 606–624, 2018.

Zhenyu Wu, Haotao Wang, Zhaowen Wang, Hailin Jin, and Zhangyang Wang. Privacy-preserving deep action recognition: An adversarial learning framework and a new dataset. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):2126–2139, 2020.

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*, pp. 22658–22668, 2023.

Zhiyuan Yan, Yuhao Luo, Siwei Lyu, Qingshan Liu, and Baoyuan Wu. Transcending forgery specificity with latent space augmentation for generalizable deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8984–8994, 2024.

Wenyuan Yang, Xiaoyu Zhou, Zhikai Chen, Bofei Guo, Zhongjie Ba, Zhihua Xia, Xiaochun Cao, and Kui Ren. Avoid-df: Audio-visual joint learning for detecting deepfake. *IEEE Transactions on Information Forensics and Security*, 18:2015–2029, 2023.

Yang Yu, Xiaolong Liu, Rongrong Ni, Siyuan Yang, Yao Zhao, and Alex C Kot. Pvass-mdd: Predictive visual-audio alignment self-supervision for multimodal deepfake detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(8):6926–6936, 2023.

Daichi Zhang, Zihao Xiao, Shikun Li, Fanzhao Lin, Jianmin Li, and Shiming Ge. Learning natural consistency representation for face forgery video detection. In *European Conference on Computer Vision*, pp. 407–424. Springer, 2024.

Xiangying Zhang, Pai Zheng, Tao Peng, Dai Li, Xujun Zhang, and Renzhong Tang. Privacy-preserving activity recognition using multimodal sensors in smart office. *Future Generation Computer Systems*, 148:27–38, 2023.

Zhixiang Zhang, Thomas Cilloni, Charles Walter, and Charles Fleming. Multi-scale, class-generic, privacy-preserving video. *Electronics*, 10(10):1172, 2021.

Hanqing Zhao, Wenbo Zhou, Dongdong Chen, Tianyi Wei, Weiming Zhang, and Nenghai Yu. Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 2185–2194, 2021.

Yinglin Zheng, Jianmin Bao, Dong Chen, Ming Zeng, and Fang Wen. Exploring temporal coherence for more general video face forgery detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 15044–15054, 2021.

Yipin Zhou and Ser-Nam Lim. Joint audio-visual deepfake detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 14800–14809, 2021.

Heqing Zou, Meng Shen, Yuchen Hu, Chen Chen, Eng Siong Chng, and Deepu Rajan. Cross-modality and within-modality regularization for audio-visual deepfake detection. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4900–4904. IEEE, 2024.

A Appendix

A.1 Related Works

Visual only Deepfake Detection Most visual-only deepfake detection methods focus on the facial region, targeting artifacts introduced during manipulation. Some approaches exploit temporal (Choi et al., 2024; Zheng et al., 2021) or spatiotemporal (Wang et al., 2023; Zhang et al., 2024) inconsistencies, while others detect blending (Li et al., 2020) or texture artifacts (Liu et al., 2020; Zhao et al., 2021). Frequency-based techniques (Liu et al., 2024; Wu et al., 2023), identity-driven representations (Cozzolino et al., 2021; Huang et al., 2023), and semantic and noise cues (Chen et al., 2024; Qiao et al., 2024; Xu et al., 2023) have also been explored for visual forgery detection. Recent works (Yan et al., 2024; Tan et al., 2024; Lin et al., 2024) have since emphasized cross-dataset robustness. LipForensics (Haliassos et al., 2021) improves generalization by analyzing lip motion. RealForensics (Haliassos et al., 2022) leverages audio-visual pretraining but ultimately relies on visual cues for classification. Further improvements (Wang et al., 2022; Prashnani et al., 2025) include adversarial learning to reduce dependency on high-frequency artifacts and phase-based analysis for enhanced temporal robustness. Despite these advances, visual-only approaches remain limited against multimodal manipulations, highlighting the need for audio-visual deepfake detection frameworks.

A.2 Implementation Details

Input Processing. Videos are extracted at 25 fps and audio is sampled at 24 kHz. Since WavTokenizer (Ji et al., 2025) reconstructs waveforms, the deepfake detector and privacy attacker convert the anonymized audio to mel-spectrograms using 64 mel filter banks, a window size of 1025, and a hop length of 240. MARLIN (Cai et al., 2023) processes video frames at 224×224; the anonymized outputs are subsequently resized to 112×112 for downstream models, following Spact (Dave et al., 2022) and STPrivacy (Li et al., 2023b).

Privacy Attributes. We evaluate five privacy attributes: gender, ethnicity, age, valence, and arousal. Gender is obtained directly from FakeAVCeleb, while ethnicity labels are obtained from VMER (Greco et al., 2020). Age is estimated using DeepFace (Serengil & Ozpinar, 2020) for video and VanPy (Koushnir et al., 2025) for audio. Video valence and arousal are estimated frame-wise using EmoNet (Toisoul et al., 2021) and averaged per video, while their audio counterparts are extracted using Wagner et al. (2023). Continuous valence and arousal values are quantized into three balanced categories (low, medium, and high) using modality-specific thresholds. We also tested age discretization with 3, 5, and 7 bins and found that 3 bins provided the best privacy supervision. This choice was verified using one-dimensional k-means clustering, resulting in similar boundaries and class proportions.

All five attributes are predicted by the unimodal audio and video outputs of the privacy attacker. For the multimodal output, we use gender and ethnicity, which are directly available or consistently annotated across modalities, avoiding modality-specific noise introduced by pretrained attribute estimators.

Model Configuration. We use $N = 3$ FSB sub-blocks. The video FSB employs depthwise 3D convolutions with kernel size $(3, 5, 5)$, while the audio FSB uses depthwise 1D convolutions with kernel size 5. Since the video decoder dimension is half the encoder dimension, the geometric ratio is $r = (1/2)^{1/3}$. Ablations over the number of FSB sub-blocks and convolution kernel sizes are reported in Sec. A.4.8.

We use NFFP (Anshul et al., 2025b) as the primary audio-visual deepfake detector, with $L_{\mathrm{DFD}} = L_{\mathrm{BCE}} + L_{\mathrm{con}}$, where L_{con} is averaged across the audio, video, and audio-visual representations. The privacy attacker is adapted from AV-HuBERT (Shi et al., 2022), using a ConvTransformer video encoder, a ViT-Tiny (Dosovitskiy, 2021) audio encoder, a cross-modal Transformer encoder, and attribute-specific MLP classification heads. The unimodal attribute classifiers are applied directly to their respective audio and video encoder outputs, while the multimodal attribute classifiers operate on the output of the cross-modal Transformer encoder.

Training Configuration. Initialization and adversarial training use 64-frame clips to balance temporal context and computational cost. Privacy evaluation also uses 64 frames, while deepfake detection uses 256 frames to provide richer temporal context. Initialization is performed for up to 30 epochs with early stopping (patience 5), adversarial training for 15 epochs, and final evaluation for 10 epochs. All models are optimized using AdamW with a learning rate of 10^{-4}. For L_{Privacy}, the modality weights w_i control the contribution of the audio, video, and audio-visual privacy losses. We use $(w_a, w_v, w_{av}) = (1, 2, 1)$, which yielded strong privacy suppression. Adversarial training uses a CosineAnnealingLR scheduler, while initialization and evaluation use ReduceLROnPlateau following the corresponding detector and privacy-attacker implementations. The FSB mixing coefficient λ follows a smooth $\tanh$ schedule from 0 to 1, reaching $\lambda = 1$ after two-thirds of adversarial training. The decoder then operates entirely on the FSB outputs for the remaining training steps.

SPLAVU Implementation: Since SPLAVU performs latent-level anonymization using frozen encoders without decoders, we adapt it to our deepfake detection setting and extend it to the audio modality. For a fair comparison, we use MARLIN and WavTokenizer as the frozen video and audio encoders, respectively. Following SPLAVU, the encoder outputs are processed by its Anonymizing Adapter Modules (AAMs) and then passed to the NFFP deepfake detector. Since NFFP originally uses its own encoders trained from scratch, we replace them with convolutional layers that align the AAM outputs with the NFFP feature dimensions. Training closely follows the original SPLAVU framework: we first initialize the adapted NFFP for deepfake detection and the AAMs as identity mappings, followed by anonymization training using the detection loss, self-supervised NT-Xent budget privacy loss, and latent consistency loss. Finally, we freeze the learned AAMs and retrain the deepfake detector and privacy attacker from scratch on their outputs for evaluation.

A.3 DATASETS

We train on **FakeAVCeleb** (Khalid et al., 2021), augmenting its real samples with **VoxCeleb2** (Chung et al., 2018) to balance the real and fake classes. **KoDF** (Kwon et al., 2021) is used for cross-dataset evaluation, while **VMER** (Greco et al., 2020) serves as an auxiliary dataset for obtaining gender and ethnicity annotations.

FakeAVCeleb (Khalid et al., 2021). FakeAVCeleb is an audio-visual deepfake dataset containing 21,566 video clips, including 500 authentic videos from VoxCeleb2 (Chung et al., 2018) and deepfakes generated using FaceSwapGAN (Nirkin et al., 2019), SV2TTS (Jia et al., 2018), FaceSwap (Korshunova et al., 2017), and Wav2Lip (Prajwal et al., 2020). The samples are organized into four categories: Real Visual-Real Audio, Real Visual-Fake Audio, Fake Visual-Real Audio, and Fake Visual-Fake Audio. To reduce the strong real-fake imbalance, we augment the real class with approximately 18,000 VoxCeleb2 samples, resulting in 18,500 real and 21,566 fake samples used for training. The manipulations applied to the real split of FakeAVCeleb ensure that the source and target identities have the same gender and ethnicity, thereby maintaining the consistency of the privacy labels in our setting.

VoxCeleb2 (Chung et al., 2018). VoxCeleb2 is a large-scale audio-visual dataset containing over one million utterances from more than 6,000 speakers collected from YouTube videos. We use approximately 18,000 samples to augment the real class of FakeAVCeleb. VoxCeleb2 is also used to initialize the SPAct (Dave et al., 2022) anonymizers for reconstruction. Gender and ethnicity annotations for the selected VoxCeleb2 samples are obtained through the VMER identity mapping described below.

KoDF (Kwon et al., 2021). KoDF is a large-scale deepfake dataset comprising 403 Korean subjects, with over 62,000 real videos and more than 175,000 deepfakes generated using six synthesis techniques. We use a subset of KoDF exclusively for evaluating cross-dataset detection generalization; the PPDD anonymizers and detector are not trained or adapted on KoDF.

VMER (Greco et al., 2020). VMER contains 3,309,742 face images from 9,129 VGGFace2 identities (Cao et al., 2018), annotated with gender and ethnicity using multiple annotators and majority voting. We use VMER only as an auxiliary annotation source. Specifically, VMER provides gender and ethnicity labels associated with VGGFace2 identities, while VoxCeleb2 provides identity

Table 4: **Comparison with modality-specific privacy-preserving methods on FakeAVCeleb.** Best results within each modality are **bold**.

Method	Modality	Acc↑	AUC↑	cMAP↓
DPRP (Nken et al., 2025)	Video	72.20	86.30	**33.40**
PPDD (Ours)	Video	**95.25**	**96.24**	35.21
CIO (Li et al., 2026)	Audio	96.90	99.10	81.10
PPDD (Ours)	Audio	**99.10**	**99.50**	**56.95**

Table 5: **Cross-dataset privacy generalization on KoDF.** Privacy attackers trained on FakeAVCeleb representations are evaluated directly on KoDF. Lower cMAP indicates lower privacy leakage.

Attributes	Representation	cMAP↓		
		Audio	Video	AV
All	Raw	44.29	45.47	46.21
	PPDD	**33.67**	**35.65**	**34.76**
w/o Ethnicity	Raw	51.30	52.91	88.62
	PPDD	**36.82**	**39.53**	**54.28**
Gender Only	Raw	87.01	87.85	88.62
	PPDD	**48.08**	**65.96**	**54.28**

correspondences to VGGFace2. We use this shared identity mapping to transfer the VMER gender and ethnicity annotations to the corresponding VoxCeleb2 samples, including those contained in FakeAVCeleb and those added for real-class augmentation.

A.4 Additional Studies

A.4.1 Comparison with Modality-Specific Privacy-Preserving Methods.

We further compare PPDD with recent modality-specific privacy-preserving approaches. For video, we adapt DPRP (Nken et al., 2025), a differential-privacy-based framework originally proposed for human activity recognition, to deepfake detection. For audio, we compare with CIO (Li et al., 2026), which uses residual vector quantization to remove semantic content while preserving acoustic information relevant to audio deepfake detection. As shown in Table 4, DPRP achieves slightly lower video privacy leakage, but with a substantial reduction in detection utility. In contrast, PPDD retains substantially higher detection performance with comparable privacy suppression. For audio, PPDD improves both detection utility and privacy preservation over CIO. These results show that PPDD provides a strong privacy-utility trade-off for each modality independently, despite being jointly learned in the audio-visual setting.

A.4.2 Cross-Dataset Privacy Generalization.

To evaluate whether the learned privacy suppression generalizes across datasets, we use a 200-sample subset of KoDF (Kwon et al., 2021) and manually annotate gender for each video. Age, valence, and arousal are obtained using the pretrained attribute models described in Sec. A.2. We then evaluate privacy attackers trained on the raw and PPDD representations of FakeAVCeleb (Khalid et al., 2021) directly on the corresponding KoDF representations, without cross-dataset adaptation.

We consider three attribute settings to account for differences in attribute availability and distribution between FakeAVCeleb and KoDF. Since KoDF contains Korean subjects, ethnicity collapses to a single class in the evaluated subset. We therefore evaluate both the full attribute set and a setting without ethnicity. In the latter, the audio and video attackers still predict gender, age, valence, and arousal, whereas the multimodal attacker predicts only gender. Because age, valence, and arousal are

Table 6: **Transferability of PPDD representations to different deepfake detectors.** We compare detection performance using raw inputs and the learned privacy-preserving representations.

Detector	Modality	Input	Acc↑	AP↑
ConvTransformer	Video	Raw	92.98	96.82
		PPDD	**94.91**	**97.75**
ViT-Tiny	Audio	Raw	**99.70**	**99.97**
		PPDD	**99.70**	99.96
MRDF (Zou et al., 2024)	Audio-Visual	Raw	**97.25**	98.60
		PPDD	94.83	**99.37**

obtained from pretrained estimators and may exhibit dataset-specific distributions, we additionally report a gender-only setting for all modalities.

As shown in Tab. 5, PPDD representations consistently reduce privacy leakage across all three evaluation settings. In the gender-only setting, cMAP decreases from 87.01 to 48.08 for audio, 87.85 to 65.96 for video, and 88.62 to 54.28 for audio-visual representations. Similar reductions are observed when additional privacy attributes are included. These results show that privacy suppression learned on FakeAVCeleb remains effective under the shift to KoDF, providing evidence that the privacy-preserving representations generalize beyond the dataset used during adversarial training.

A.4.3 Unimodal Deepfake Detection.

We first examine whether the learned privacy-preserving representations retain modality-specific manipulation cues independently of multimodal fusion. We train separate deepfake detectors on the audio and video modalities, using a ConvTransformer for video and ViT-Tiny for audio. As shown in Tab. 6, detection performance is preserved when the raw inputs are replaced with PPDD representations for both modalities. In particular, the audio detector maintains nearly identical performance, while the video detector improves across the reported metrics. These results indicate that the learned representations retain useful modality-specific manipulation information and do not rely on complementary information from the other modality to support deepfake detection.

A.4.4 Generalization to Different Deepfake Detectors.

We further examine whether the learned representations are specific to the deepfake detector used during adversarial training. In addition to two unimodal detectors discussed in Sec. A.4.3, we evaluate the detection generalization on a different multimodal deepfake detector. We replace NFFP with MRDF (Zou et al., 2024) and train the detector independently using either raw inputs or the frozen PPDD representations. As shown in Tab. 6, MRDF retains comparable detection performance on the PPDD representations Together with the unimodal results, this shows that the learned representations are not tied to the detector used during PPDD training and can support different deepfake detection architectures.

A.4.5 Impact of Multimodal Privacy Loss during Anonymization

Our framework uses both unimodal and multimodal privacy losses to suppress privacy information available within individual modalities and through complementary audio-visual cues. To assess the contribution of multimodal supervision, we remove the multimodal privacy loss and train the anonymizers using only the unimodal privacy outputs. We then evaluate the resulting representations using the original privacy attacker. Removing the multimodal loss increases the average cMAP from 41.96 to 43.91. This shows that multimodal privacy supervision provides additional suppression of privacy information that can otherwise remain recoverable across the two modalities.

A.4.6 Impact of the Geometric Ratio r.

The geometric ratio $r = (d_v^{\text{dec}}/d_v)^{1/N}$ affects the FSB in two ways. First, each video FSB sub-block progressively reduces the feature dimension by r, such that the representation reaches the decoder

Table 7: **Impact of the geometric ratio r.** PD denotes progressive down-projection across FSB sub-blocks. The complete model uses r for both progressive projection and suppression regularization.

PD	L_{Suppress}	AUC↑	cMAP↓			
			Audio	Video	AV	Avg.
×	×	93.99	79.53	84.82	38.34	67.56
✓	×	95.13	61.80	**33.20**	33.83	42.94
×	✓	95.26	61.09	40.82	34.05	45.32
✓	✓	96.24	**56.95**	35.21	**33.71**	**41.96**

dimension after N sub-blocks. This progressive down-projection enables smooth and consistent suppression of privacy-related features across all FSB stages under the guidance of the suppression loss. Second, the same ratio is used by L_{Suppress} to control the amount of feature suppression at each sub-block. Thus, r jointly governs progressive feature projection and controlled suppression across the FSB stages.

To comprehensively study the impact of r, we evaluate all four combinations of progressive down-projection (PD) and suppression loss. When both are disabled, neither mechanism is guided by r, corresponding to the effective $r = 1$ setting. Enabling either component independently isolates the two roles of r, while enabling both corresponds to our complete configuration.

As shown in Tab. 7, disabling both mechanisms results in substantially higher privacy leakage, with an average cMAP of 67.56, while also reducing detection AUC to 93.99. Introducing progressive down-projection alone substantially reduces the average cMAP to 42.94 and improves AUC to 95.13. In comparison, applying L_{Suppress} without progressive down-projection achieves an average cMAP of 45.32 and an AUC of 95.26. These results show that both roles of r independently contribute to the privacy-utility trade-off. In particular, gradual down-projection suppresses privacy information more effectively than performing the dimensional projection in a single step after the FSB.

Combining both mechanisms provides the strongest overall performance, achieving the lowest average cMAP of 41.96 and the highest AUC of 96.24. The two roles of r therefore complement each other: progressive down-projection gradually transforms the representation toward the decoder space, while L_{Suppress} controls the amount of information suppressed at each stage. Their combination enables ratio-guided suppression throughout the FSB and yields the strongest overall privacy-utility trade-off.

A.4.7 Impact of λ Schedule.

We further study the effect of progressively introducing the FSB during adversarial representation learning. Our default setting uses the $\tanh$-based schedule $\lambda(t)$ to gradually transition from the pretrained projection pathway to the FSB, reaching $\lambda = 1$ after two-thirds of anonymizer training. We compare this with setting $\lambda = 1$ throughout training, such that the representations pass entirely through the FSB from the beginning. Without the progressive schedule, detection AUC decreases from 96.24 to 95.11, while average privacy cMAP increases from 41.96 to 42.74. This shows that gradually introducing the FSB provides a smoother transition from the pretrained representations, improving both privacy preservation and detection utility.

A.4.8 Hyperparameter Ablations.

Since the FSB is central to learning the privacy-preserving representations, we analyze two of its key hyperparameters: the kernel size of the depthwise convolution and the number of suppression sub-blocks N. We evaluate kernel sizes $\{1, 3, 5, 7\}$ with $N = 3$, and $N \in \{2, 3, 4\}$ with kernel size 5. Following the video encoder design (Cai et al., 2023), the reshaped video representation has spatial dimensions 14×14; we therefore restrict the maximum kernel size to 7 to avoid excessive boundary padding.

As shown in Tab. 8, kernel size 5 with $N = 3$ achieves the highest detection AUC and lowest average privacy cMAP, and is therefore used in our final configuration. Smaller kernels provide less local context for estimating the suppression maps, while increasing the kernel to 7 substantially degrades

Table 8: **FSB hyperparameter ablations.** We evaluate the depthwise convolution kernel size and the number of Feature Suppression sub-blocks (N).

Kernel	N	AUC↑	cMAP↓			
			Audio	Video	AV	Avg.
1	3	92.61	63.86	41.71	36.28	47.28
3	3	94.37	57.84	40.70	34.21	44.25
5	**3**	**96.24**	**56.95**	**35.21**	**33.71**	**41.96**
7	3	94.24	60.65	90.42	36.93	62.67
5	2	94.85	82.43	33.89	34.65	50.32
5	**3**	**96.24**	**56.95**	35.21	**33.71**	**41.96**
5	4	92.85	99.81	89.31	48.32	79.15

Table 9: **Age-bin selection.** Privacy-attacker performance on raw representations using different numbers of age bins. Higher cMAP indicates a stronger privacy attacker before anonymization.

Age Bins	Audio cMAP↑	Video cMAP↑
3	**92.84**	**88.19**
5	91.62	79.32
7	89.28	71.21

video privacy suppression. Similarly, $N = 3$ provides a better privacy–utility trade-off than using two or four suppression stages. These results suggest that moderate local context and progressive suppression across three stages provide a suitable configuration for learning the privacy-preserving representations.

A.4.9 Age-Bin Selection.

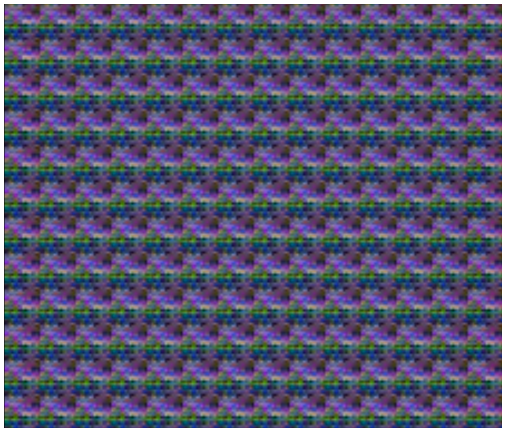
As described in the main paper, continuous age, valence, and arousal values are quantized into three categories using modality-specific thresholds to maintain balanced distributions. To determine the age discretization, we evaluate privacy attackers trained with 3, 5, and 7 age bins on the raw representations. Since age is used only by the audio and video privacy outputs, we report cMAP for these two modalities.

As shown in Tab. 9, three age bins yield the highest privacy-attacker performance for both modalities, while performance decreases as the number of bins increases. We therefore use three bins throughout our experiments. Importantly, this choice provides the anonymizers with the strongest privacy attacker during adversarial training, allowing privacy suppression to be learned against the most effective age-inference setting among those evaluated. The decrease with additional bins may result from the finer boundaries between adjacent age categories, which make age inference increasingly difficult.

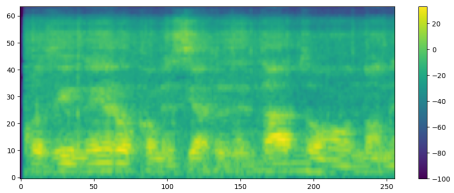
A.4.10 Qualitative Analysis

We qualitatively analyze the anonymized outputs produced by PPDD using 100 randomly selected FakeAVCeleb samples, distributed across the considered privacy attributes. An example anonymized video frame and the mel-spectrograms of the corresponding raw and anonymized audio are shown in Fig. 3. Additional examples for both real and deepfake samples are provided in the supplementary files.

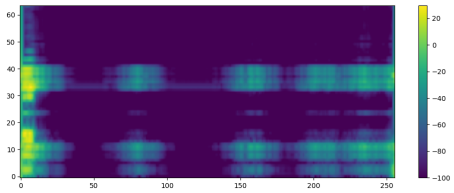
As shown in Fig. 3a, the output of the video anonymizer no longer preserves a human-interpretable facial appearance. The anonymized frame is highly transformed and the identity of the person cannot be directly interpreted from it. Within each anonymized video, the visual appearance also remains largely similar across frames. Across different input videos, the anonymized outputs contain limited visually distinguishable identity-specific information, making it difficult to differentiate the original identities through direct visual inspection.



(a) Privacy-preserving visual representation



(b) Raw audio mel-spectrogram



(c) Privacy-preserving audio representation

Figure 3: **Example privacy-preserving representations.** (a) Visual representation produced by the video anonymizer. (b) Mel-spectrogram of the corresponding raw audio. (c) Mel-spectrogram of the privacy-preserving audio representation.

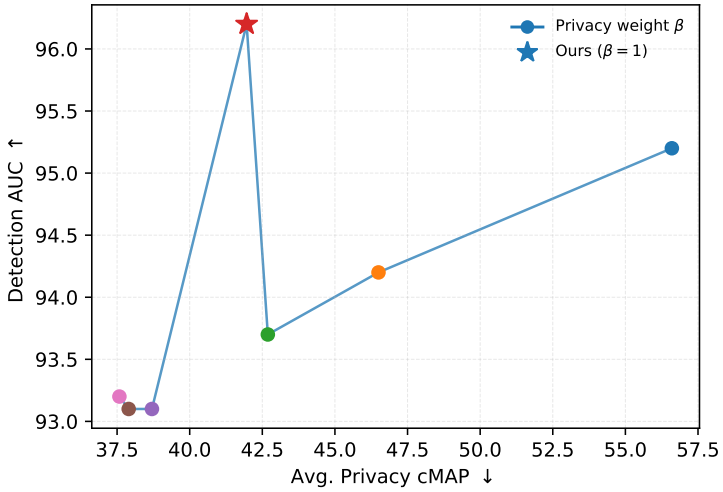


Figure 4: **Privacy-utility trade-off.** Detection AUC versus average privacy cMAP obtained by varying the privacy-loss weight β during anonymization. Lower cMAP indicates stronger privacy preservation, while higher AUC indicates better deepfake detection performance.

A similar effect can be observed for audio. Comparing Fig. 3b and Fig. 3c, the anonymized audio shows substantial suppression of the information visible in the raw mel-spectrogram. While the raw audio contains rich spectral patterns across time and frequency, the anonymized output is considerably sparser and retains only a reduced spectral structure. Similar suppression is observed across the sampled audio outputs.

Overall, these examples illustrate the representation-learning objective of PPDD. The anonymizers produce outputs in the original video and audio signal formats, allowing them to be directly processed by downstream deepfake detectors, while the anonymized outputs are no longer human-interpretable in the same way as the original inputs. Together with the quantitative results, this shows that the learned representations can support strong deepfake detection while substantially reducing information available for privacy-attribute inference.

A.4.11 PRIVACY-UTILITY TRADE-OFF.

We investigate how the strength of privacy suppression affects the trade-off between privacy preservation and deepfake detection utility. To control this trade-off, we introduce a privacy-loss weight β during the anonymizer update, such that the anonymizer optimizes

$$L_{\mathrm{A}} = L_{\mathrm{Anony}} + L_{\mathrm{DFD}} - \beta L_{\mathrm{Privacy}}. \tag{13}$$

We vary β while keeping the remaining loss weights fixed and plot the resulting detection AUC against average privacy cMAP in Fig. 4. Overall, stronger privacy weighting leads to lower privacy leakage, although the relationship is not strictly monotonic due to the adversarial optimization. We select $\beta = 1$ for our final model, as it achieves the highest detection AUC among the evaluated settings while providing substantial privacy suppression. Increasing β beyond this operating point further reduces privacy leakage, but with a decrease in detection performance. Thus, larger values of β can be selected when stronger privacy preservation is preferred, and additional detection degradation is acceptable.

A.5 TRAINING ALGORITHM

Algorithm 1 Training and Evaluation of PPDD

Input: Dataset $\mathcal{D} = \{(v, a, y, \mathbf{s})\}$
Anonymizers $\mathrm{A}_a = \{\mathrm{E}_a, \mathrm{F}_a, \mathrm{Q}_a, \mathrm{G}_a\}$, $\mathrm{A}_v = \{\mathrm{E}_v, \mathrm{F}_v, \mathrm{Q}_v, \mathrm{G}_v\}$
Deepfake detector D; Privacy attacker P
Number of steps: $T_{\mathrm{A}}, T_{\mathrm{P}}, T_{\mathrm{D}}$
Learning rates: $\eta_{\mathrm{A}}, \eta_{\mathrm{P}}, \eta_{\mathrm{D}}$
Output: $\theta_{\mathrm{A}_a}, \theta_{\mathrm{A}_v}, \theta_{\mathrm{D}}, \theta_{\mathrm{P}}$

Stage I: Initialization

Initialize $\theta_{\mathrm{E}_a}, \theta_{\mathrm{Q}_a}, \theta_{\mathrm{G}_a}$ with WavTokenizer (Ji et al., 2025)
Initialize $\theta_{\mathrm{E}_v}, \theta_{\mathrm{Q}_V}, \theta_{\mathrm{G}_v}$ with Marlin (Cai et al., 2023)
Initialize $\theta_{\mathrm{F}_a}, \theta_{\mathrm{F}_v}$ from scratch
for $t \leftarrow 1$ **to** T_{D} **do**
    $\theta_{\mathrm{D}} \leftarrow \theta_{\mathrm{D}} - \eta_{\mathrm{D}} \nabla_{\theta_{\mathrm{D}}} L_{\mathrm{DFD}}$
end for
for $t \leftarrow 1$ **to** T_{P} **do**
    $\theta_{\mathrm{P}} \leftarrow \theta_{\mathrm{P}} - \eta_{\mathrm{P}} \nabla_{\theta_{\mathrm{P}}} L_{\mathrm{Privacy}}$
end for

Stage II: Adversarial Representation Learning

for $t \leftarrow 1$ **to** T_{A} **do**
    $z_m = E_m(m)$, for modality $m \in \{a, v\}$
    $\hat{z}_m = F_m(z_m)$, for modality $m \in \{a, v\}$
    $\lambda(t) \leftarrow \mathrm{TanhSchedule}(t, T_{\mathrm{A}})$
    $z_m^{\lambda} = \lambda(t)\hat{z}_m + (1 - \lambda(t))Q_m(z_m)$, for modality $m \in \{a, v\}$
    $\tilde{m} = G_m(z_m^{\lambda})$, for modality $m \in \{a, v\}$

    if $t \bmod 2 = 1$ **then**
    // Anonymizer update
        $\theta_{\mathrm{A}_a} \leftarrow \theta_{\mathrm{A}_a} - \eta_{\mathrm{A}} \nabla_{\theta_{\mathrm{A}_a}} (L_{\mathrm{Anony}} + L_{\mathrm{DFD}} - L_{\mathrm{Privacy}})$
        $\theta_{\mathrm{A}_v} \leftarrow \theta_{\mathrm{A}_v} - \eta_{\mathrm{A}} \nabla_{\theta_{\mathrm{A}_v}} (L_{\mathrm{Anony}} + L_{\mathrm{DFD}} - L_{\mathrm{Privacy}})$
    else
    // Detector and privacy-attacker update
        $\theta_{\mathrm{D}} \leftarrow \theta_{\mathrm{D}} - \eta_{\mathrm{D}} \nabla_{\theta_{\mathrm{D}}} L_{\mathrm{DFD}}$
        $\theta_{\mathrm{P}} \leftarrow \theta_{\mathrm{P}} - \eta_{\mathrm{P}} \nabla_{\theta_{\mathrm{P}}} L_{\mathrm{Privacy}}$
    end if
end for

Stage III: Downstream Evaluation

Freeze $\theta_{\mathrm{A}_a}, \theta_{\mathrm{A}_v}$
Reinitialize $\theta_{\mathrm{D}}, \theta_{\mathrm{P}}$
for $t \leftarrow 1$ **to** T_{D} **do**
    $\tilde{a} = A_a(a)$; $\tilde{v} = A_v(v)$
    $L_{\mathrm{DFD}} \leftarrow L_{\mathrm{det}}\big(D(\tilde{v}, \tilde{a}), y\big)$
    $\theta_{\mathrm{D}} \leftarrow \theta_{\mathrm{D}} - \eta_{\mathrm{D}} \nabla_{\theta_{\mathrm{D}}} L_{\mathrm{DFD}}$
end for
for $t \leftarrow 1$ **to** T_{P} **do**
    $\tilde{a} = A_a(a)$; $\tilde{v} = A_v(v)$
    $L_{\mathrm{Privacy}} \leftarrow L_{\mathrm{priv}}\big(P(\tilde{v}, \tilde{a}), \mathbf{s}\big)$
    $\theta_{\mathrm{P}} \leftarrow \theta_{\mathrm{P}} - \eta_{\mathrm{P}} \nabla_{\theta_{\mathrm{P}}} L_{\mathrm{Privacy}}$
end for