



# Cross-Modal Watermarking for Authentic Audio Recovery and Tamper Localization in Synthesized Audiovisual Forgeries

Minyoung Kim<sup>1</sup>, Sehwan Park<sup>1</sup>, Sungmin Cha<sup>2</sup>, Paul Hongsuck Seo<sup>1</sup>

<sup>1</sup>Dept. of CSE, Korea University, Republic of Korea

<sup>2</sup>New York University, USA

{omniverse186, shp216, phseo}@korea.ac.kr

sungmin.cha@nyu.edu

## Abstract

Recent advances in voice cloning and lip synchronization models have enabled Synthesized Audiovisual Forgeries (SAVFs), where both audio and visuals are manipulated to mimic a target speaker. This significantly increases the risk of misinformation by making fake content seem real. To address this issue, existing methods detect or localize manipulations but cannot recover the authentic audio that conveys the semantic content of the message. This limitation reduces their effectiveness in combating audiovisual misinformation. In this work, we introduce the task of Authentic Audio Recovery (AAR) and Tamper Localization in Audio (TLA) from SAVFs and propose a cross-modal watermarking framework to embed authentic audio into visuals before manipulation. This enables AAR, TLA, and a robust defense against misinformation. Extensive experiments demonstrate the strong performance of our method in AAR and TLA against various manipulations, including voice cloning and lip synchronization.<sup>1</sup>

**Index Terms:** Synthesized Audiovisual Forgeries, Authentic Audio Recovery, Tamper Localization in Audio

## 1. Introduction

Recent advancements in generative speech models [1, 2, 3] have enabled the synthesis of high-fidelity audio content that closely resembles real-world speech. Among these, voice cloning techniques [4, 5, 6] can replicate a speaker’s unique vocal characteristics from just a few audio samples, facilitating personalized content generation. When combined with lip synchronization methods [7, 8, 9] that generate photorealistic video sequences aligned with input speech, these technologies enable the creation of highly realistic audiovisual content. Such advancements have broad applications in digital media, virtual avatars, language dubbing, and assistive technologies, where scalable, high-quality audio-visual synchronization is crucial.

However, these techniques also pose significant risks, particularly in the spread of highly realistic misinformation. Especially, Synthesized Audiovisual Forgeries (SAVFs)—videos in which both the audio and visuals have been manipulated through voice cloning and lip synchronization—can be exploited to impersonate individuals, manipulate public opinion, and undermine trust in digital media [10, 11]. To mitigate these risks, efforts have focused on detecting and localizing fake visual or audio content in SAVFs. Note that one approach is the localization of tampered regions [12, 13], which can help in understanding the attacker’s intent and may enable the partial reuse of manipulated content. While localization provides

valuable insights, it remains insufficient for assessing the significance of the altered regions or determining the extent of semantic shifts compared to the authentic audio. As a result, it has limitations in fully capturing the attacker’s intent and does not allow for the complete reuse of the content.

In this work, we propose a novel task of authentic audio recovery from SAVFs to address the limitations of existing approaches. Instead of merely detecting or localizing manipulated content, our goal is to reconstruct the authentic audio signal, which directly conveys the semantic content of the message. To achieve this, we introduce a watermarking-based approach that embeds the authentic audio into visual frames before any potential forgery. This cross-modal approach enables the recovery of authentic audio even when the audio is partially or entirely removed during the forgery process, where direct restoration is particularly challenging. Beyond audio reconstruction, our method also aids in localizing tampered regions by detecting manipulated content from recovered audio, as manually comparing and identifying altered parts can be highly laborious.

Through extensive experiments, we show that our approach enables the robust recovery of authentic audio, even when the audio stream is altered or replaced. Furthermore, it precisely localizes tampered regions, providing a proactive defense against audiovisual misinformation. Notably, our approach remains robust even when trained on datasets without human faces or voices, addressing privacy and portrait rights concerns. The core contributions of this paper can be summarized as follows:

- We introduce the novel task of recovering authentic audio from SAVFs, moving beyond detection and localization to restore authentic audio content.
- We propose a cross-modal watermarking framework that embeds audio into visual frames, ensuring robust tamper localization and audio recovery even after various manipulations.
- Experimental results demonstrate that our method enables robust recovery of authentic audio and extends beyond human faces and voices.

## 2. Related Works

**Voice Cloned Audio Localization** Recent advances in voice cloning have intensified the challenge of localizing manipulated audio segments. Approaches like BAM [14] and CFPRF [15] utilize boundary-aware attention and coarse-to-fine refinement to detect tampering; however, their reliance on specific training manipulations limits robustness against novel attacks. Proactive watermarking methods such as Wavmark [16] and Audioseal [13] verify embedded watermarks to identify altered regions, yet they fall short of recovering the authentic audio.

**INN-based Steganography and Watermarking** Steganog-

<sup>1</sup>The code is available at: [https://eurominyoung186.github.io/CMW\\_SAVF/](https://eurominyoung186.github.io/CMW_SAVF/)

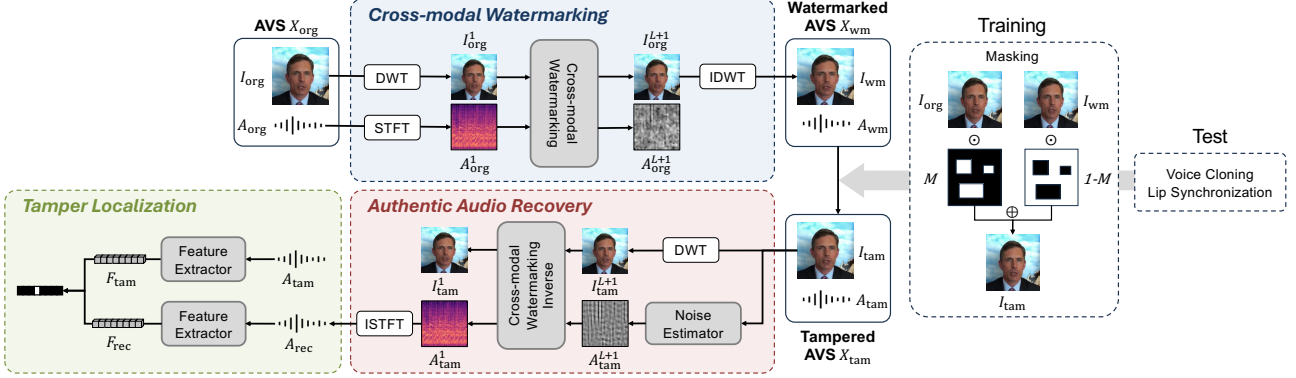


Figure 1: **Overall architecture of our model.** The framework comprises three main processes: cross-modal watermarking (CMW), authentic audio recovery, and tamper localization. In the CMW process, CMW embed the authentic audio within a visual frame. For authentic audio recovery, noise estimators predict the transformed audio output from watermarked visual frame, enabling the inverse CMW to recover the authentic audio embedded in the visual frame. Finally, in tamper localization, we compute feature maps for both the recovered and tampered audio to generate a score that identifies the tampered regions.

raphy and watermarking embed data within cover content for secure or traceable transfer. Traditional techniques [17, 18] are constrained by capacity and invisibility issues, leading to the emergence of deep learning methods like [19]. Invertible Neural Networks (INNs) [20] offer precise embedding and extraction, as evidenced by HiNet [21], LF-VSN [22] and ThinImg [23]—the latter hiding audio within images via mel-spectrograms.

### 3. Method

#### 3.1. Overview

In this work, we particularly focus on SAVFs utilized to manipulate the original message of the speaker by modifying real speech videos through voice cloning and lip synchronization. Our task prioritizes Authentic Audio Recovery (AAR) and Tamper Localization in Audio (TLA) because speech conveys core semantic content, serving as the primary medium for delivering factual and persuasive information. Formally, our goal is twofold: (1) to recover the authentic audio signal from tampered audiovisual stream and (2) to identify time intervals  $\{(t_{\text{start}}^i, t_{\text{end}}^i)\}_{i=1}^N$  that correspond to tampered audio regions where each pair  $(t_{\text{start}}^i, t_{\text{end}}^i)$  represents a timestamp range indicating a segment of containing tampered content.

To address the tasks of AAR and TLA, we propose a watermarking-based approach, as illustrated in Fig. 1. For notational simplicity, we assume that video  $V$  consists of a single visual frame  $I$  and the audio segment  $A$  that occurs during its display interval. It is important to note that our technique, formulated under this simplified setting, can be seamlessly extended to Audiovisual Stream (AVS) of arbitrary length by segmenting them into corresponding visual frames and audio segments. Specifically, given an original AVS  $X_{\text{org}} = (I_{\text{org}}, A_{\text{org}})$  as input, where  $I_{\text{org}}$  is visual frame and  $A_{\text{org}}$  is its corresponding authentic audio,  $A_{\text{org}}$  is imperceptibly embedded into the visual frame  $I_{\text{org}}$ , producing a watermarked AVS  $X_{\text{wm}} = (I_{\text{wm}}, A_{\text{wm}})$ . In this process,  $I_{\text{wm}}$  contains information of  $A_{\text{org}}$ , while the audio remains unchanged, indicating that  $A_{\text{wm}}$  is identical to  $A_{\text{org}}$ . This watermarked AVS  $X_{\text{wm}}$  can be tampered with methods like lip synchronization and voice cloning, resulting in a tampered AVS  $X_{\text{tam}} = (I_{\text{tam}}, A_{\text{tam}})$ . For AAR and TLA, we estimate  $A_{\text{rec}}$ , a recovered version of  $A_{\text{org}}$ , from  $I_{\text{tam}}$ . By comparing  $A_{\text{rec}}$  and  $A_{\text{tam}}$ , we can localize tampered region. The detailed method for both processes will be elaborated in the following sections.

#### 3.2. Audio Embedding with Cross-Modal Watermarking

In our proposed approach, we embed the authentic audio signal  $A_{\text{org}}$  into the visual frame  $I_{\text{org}}$  using Cross-Modal Watermarking (CMW). To embed  $A_{\text{org}}$  as a watermark into  $I_{\text{org}}$ , we adopt trainable Invertible Neural Network (INN) blocks [22] with the unique property of reversibility, allowing for exact recovery of inputs from outputs.

Formally, given an AVS with a single frame  $X_{\text{org}} = (I_{\text{org}}, A_{\text{org}})$ , our cross-modal watermarking module constructs a watermarked AVS  $X_{\text{wm}} = (I_{\text{wm}}, A_{\text{wm}})$ . An INN block at layer  $l$  takes inputs  $I_{\text{org}}^l$  and  $A_{\text{org}}^l$  and produces  $I_{\text{org}}^{l+1}$  and  $A_{\text{org}}^{l+1}$  as:

$$I_{\text{org}}^{l+1} = I_{\text{org}}^l + \phi(A_{\text{org}}^l), \quad (1)$$

$$A_{\text{org}}^{l+1} = A_{\text{org}}^l \odot \exp(\sigma(\rho(I_{\text{org}}^{l+1}))) + \eta(I_{\text{org}}^{l+1}), \quad (2)$$

where  $\phi$ ,  $\rho$  and  $\eta$  are neural networks,  $\sigma$  is a sigmoid activation, and  $\odot$  denotes element-wise multiplication. Note that the inversion operation does not require the direct inverses of  $\phi$ ,  $\rho$  and  $\eta$ . Instead, it relies on the ability to recover the original inputs from the outputs through the following inversion equations:

$$A_{\text{org}}^l = (A_{\text{org}}^{l+1} - \eta(I_{\text{org}}^{l+1})) \odot \exp(-\sigma(\rho(I_{\text{org}}^{l+1}))), \quad (3)$$

$$I_{\text{org}}^l = I_{\text{org}}^{l+1} - \phi(A_{\text{org}}^l). \quad (4)$$

The only constraints for this inversion are that the output shapes of  $\rho$  and  $\eta$  must match  $A_{\text{org}}^l$ , and the output shape of  $\phi$  must match  $I_{\text{org}}^l$ . These conditions ensure that the original inputs can be exactly recovered from the transformed outputs, which is essential for the watermarking operation.

In our implementation, we construct  $I_{\text{org}}^1$  by applying the Discrete Wavelet Transform (DWT) in  $I_{\text{org}} \in \mathbb{R}^{H \times W \times 3}$ , where  $H$  and  $W$  denote the height and width of visual frame. To address the shape discrepancy between  $I_{\text{org}}^1 \in \mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times 48}$  and  $A_{\text{org}} \in \mathbb{R}^S$ , where  $S$  is length of corresponding audio segment, we transform  $A_{\text{org}}$  into spectrogram by applying the Short-Time Fourier Transform (STFT) and reshape it into  $\mathbb{R}^{\frac{H}{4} \times \frac{W}{4} \times 1}$ , resulting in  $A_{\text{org}}^1$ . Then,  $I_{\text{org}}^1$  and  $A_{\text{org}}^1$  are fed into a stack of  $L$  INN blocks. This process generates outputs  $I_{\text{org}}^{L+1}$  and  $A_{\text{org}}^{L+1}$ , with  $I_{\text{wm}}$  obtained by applying Inverse DWT (IDWT) to  $I_{\text{org}}^{L+1}$ .

For  $\phi$ ,  $\rho$  and  $\eta$ , we follow the design principles in [24] using five  $3 \times 3$  convolutional layers with Leaky ReLU activation, where the output channel sizes are adjusted accordingly to match those of  $I_{\text{org}}^l$  and the reshaped  $A_{\text{org}}^l$ .

### 3.3. Authentic Audio Recovery and Tamper Localization

**Authentic Audio Recovery** Thanks to the invertible property of INN, we can perfectly recover the original inputs  $I_{\text{org}}$  and  $A_{\text{org}}$  from their outputs  $I_{\text{org}}^{L+1}$  and  $A_{\text{org}}^{L+1}$ . While  $I_{\text{org}}^{L+1}$  is accessible from the watermarked image  $I_{\text{wm}}$ ,  $A_{\text{org}}^{L+1}$  cannot be directly accessed from the watermarked AVS  $X_{\text{wm}}$ , as it is discarded during the CMW process. To overcome this, we predict  $A_{\text{org}}^{L+1}$  from  $I_{\text{wm}}$  using a noise estimator inspired by [22]. This estimation process enables the recovery of the embedded original audio  $A_{\text{org}}$  solely from the watermarked frame  $I_{\text{wm}}$  without requiring access to the authentic audio. In practice, the watermarked frame  $I_{\text{wm}}$  may also be altered within the SAVFs due to lip synchronization forgery, introducing an additional challenge. To address this, we learn a robust model capable of handling such modifications, which will be discussed later.

**Tamper Localization** After AAR, we can localize the tampered region within the SAVFs by comparing the recovered and tampered audio. However, a naïve direct comparison of raw audio signals is highly susceptible to noise and recovery errors, making it unreliable for precise tampering localization. To address this, we use a Semantic Feature Extractor (SFE) based on [25] to project the audio streams  $A$  into a semantic feature space  $F \in \mathbb{R}^{T \times C}$ . We then compute their similarity using the inner product for more robust and reliable TLA. Formally, given audio feature map  $F_{\text{tam}} = \{\mathbf{f}_{\text{tam}}^t\}$  and  $F_{\text{rec}} = \{\mathbf{f}_{\text{rec}}^t\}$  extracted from the  $A_{\text{tam}}$  and  $A_{\text{rec}}$ , we compute a cosine similarity score  $s_{\text{tam}}^t = (\mathbf{f}_{\text{tam}}^t)^\top (\mathbf{f}_{\text{rec}}^t)$  at timestep  $t$ , where  $\mathbf{f}_{\text{tam}}^t \in \mathbb{R}^C$  and  $\mathbf{f}_{\text{rec}}^t \in \mathbb{R}^C$  are temporally aligned feature vectors. This feature-level comparison enhances resilience to recovery errors and minor perturbations that do not alter the underlying semantics. We conducted experiments to evaluate the effectiveness of feature-level comparison. While a naïve direct comparison of raw audio signals yields an Average Precision (AP) of 87.17, our approach significantly improves it to 98.28, demonstrating superior robustness in tamper localization.

### 3.4. Training

We train the entire network end-to-end in an unsupervised manner, without requiring localization annotations for tampering attacks. Our total loss function is a weighted sum of four components described below: watermarking loss, visual reconstruction loss, audio reconstruction loss, and feature contrastive loss, formulated as  $\mathcal{L} = \lambda_{\text{WL}}\mathcal{L}_{\text{WL}} + \lambda_{\text{VRL}}\mathcal{L}_{\text{VRL}} + \lambda_{\text{ARL}}\mathcal{L}_{\text{ARL}} + \lambda_{\text{SFCL}}\mathcal{L}_{\text{SFCL}}$  where  $\lambda$ s are coefficients for each loss term.

**Watermarking Loss** To ensure the watermarked visual frame  $I_{\text{wm}}$  closely resembles the original visual frame  $I_{\text{org}}$ , we apply an  $L_2$  loss,  $\mathcal{L}_{\text{WL}} = \|I_{\text{wm}} - I_{\text{org}}\|_2^2$ .

**Reconstruction Losses** To ensure the original frame  $I_{\text{org}}$  and audio stream  $A_{\text{org}}$  are accurately recovered with  $A_{\text{org}}^{L+1}$  missing, we introduce two loss terms  $\mathcal{L}_{\text{VRL}} = \|I_{\text{org}} - I_{\text{rec}}\|_2^2$  and  $\mathcal{L}_{\text{ARL}} = \|A_{\text{org}} - A_{\text{rec}}\|_2^2$ . These terms are critical to train the noise estimator for  $A_{\text{org}}^{L+1}$  introduced in Section 3.2.

**Semantic Feature Contrastive Loss** To ensure robust tamper localization, we compare the tampered and recovered audio streams in a semantic feature space. Specifically, we enforce proximity between temporally aligned features  $\mathbf{f}_{\text{org}}^t$  and  $\mathbf{f}_{\text{rec}}^t$  using a contrastive loss [26] as follows:

$$\mathcal{L}_{\text{SFCL}} = \sum_t \mathcal{L}_{\text{NCE},t} = - \sum_t \log \frac{\exp(\mathbf{f}_{\text{org}}^t \cdot \mathbf{f}_{\text{rec}}^t / \tau)}{\sum_{l=1}^T \exp(\mathbf{f}_{\text{org}}^t \cdot \mathbf{f}_{\text{rec}}^l / \tau)}$$

where  $\tau$  is a temperature, and  $T$  is the number of features in the feature map.

**Masking Strategy** Lip synchronization forgery alters facial regions, removing parts of embedded watermarks and complicating AAR. To enhance robustness against such forgery, we introduce masking strategies during training that partially remove embedded watermarks in  $I_{\text{wm}}$ . As shown in Fig. 1, given an original frame  $I_{\text{org}}$  and its watermarked counterpart  $I_{\text{wm}}$ , we apply a binary mask  $M \in \mathbb{R}^{H \times W}$  to selectively replace watermarked regions using  $M \odot I_{\text{org}} + (1 - M) \odot I_{\text{wm}}$ , where  $\odot$  denotes element-wise multiplication. This simulates watermark removal, helping the model learn to recover audio even when portions are erased. Specifically, we employ two masking strategies: Random Mask Generation, which applies one to three randomly shaped geometric masks with side lengths sampled between 20 and 150 pixels, and Facial Mask Generation, which uses a facial detection model [27] to identify and alter facial regions. In inference time, we utilize lip synchronization models without applying masking strategies.

## 4. Experiments

### 4.1. Experimental Setup

**Dataset** We use the HDTF dataset [28], which consists of 410 talking face videos with synchronized speech, totaling 16 hours of audiovisual data. As one of the primary benchmarks for lip-synchronization [28, 29], HDTF provides high-quality, diverse speaker recordings, making it well-suited for evaluating the effectiveness of our approach in AAR and TLA. A random subset of 98 videos is used for training, with the remaining 312 reserved for evaluation. A random 5-second segments from each of these samples are pre-selected for fair evaluations. All videos are processed at 25 fps with audio sampled at 16 kHz.

**Evaluation Metrics** For AAR, we adopt Signal-to-Noise Ratio (SNR), Perceptual Evaluation of Speech Quality (PESQ) from [13, 16], which are common metrics, measuring the audio fidelity. For TLA, we employ Intersection over Union (IoU), Average Precision (AP), and Area Under the Curve (AUC) following [12, 30]. We additionally measure the quality of the watermarked contents using Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index Measure (SSIM) from [12].

**Implementation Details** We use six invertible blocks ( $L = 6$ ) and optimize the weights for 10K iterations using Adam optimizer with a learning rate of  $1 \times 10^{-4}$ ,  $\beta_1 = 0.9$  and  $\beta_2 = 0.5$ . For all experiments, we set  $\lambda_{\text{WL}} = 10$ ,  $\lambda_{\text{ARL}} = 10$ ,  $\lambda_{\text{VRL}} = 0.1$ ,  $\lambda_{\text{SFCL}} = 1$  and  $\tau = 0.07$  while setting the window size and hop length to 510 and 128 for STFT. The channel of the audio feature map is 32 and we employ a queue that stores 65,536 features from previous iterations.

### 4.2. Results

**Tamper Localization in Audio** Table 1 compares our model with SOTA audio tamper localization methods, including passive approaches, BAM [14] and CFPF [15] and proactive watermarking-based (WM) methods, AudioSeal [13] and Wavmark [16]. We evaluate two tampering scenarios: Audio Swapping (AS), where 10% to 30% of audio is replaced with another segment from the same speaker, and Voice Cloning (VC), where a audio segment is substituted with a synthetic voice generated by a cloning model [4]. In both cases, visual frames are also manipulated using MuseTalk [31], posing an additional challenge for our model. Unlike audio-only baselines that ignore visual inputs, our model embeds audio within the visual modality, enabling accurate recovery and tamper localization. As a result, our cross-modal watermarking model consistently out-

Table 1: **Comparison of Different Audio Tamper Localization Methods on the HDTF Dataset.** Tampering simulation uses two methods: **AS**-inserting a different audio segment into the parts of original audio and **VS**-modifying parts of the audio with voice generated by OpenVoice [4]. Localization metrics include IoU, AP, and AUC, while SNR and PESQ measure recovered audio quality, respectively. SSIM and PSNR measure video quality. **WM** and **CM** refer to watermarking and cross-modal techniques.

Name	WM	CM	Audio Recovery			Tamper Localization (AS)			Tamper Localization (VC)			Audio Quality		Visual Quality	
			SNR $\uparrow$	PESQ $\uparrow$	N/A	N/A	IoU $\uparrow$	AP $\uparrow$	AUC $\uparrow$	IoU $\uparrow$	AP $\uparrow$	AUC $\uparrow$	SNR $\uparrow$	PESQ $\uparrow$	PSNR $\uparrow$
CFPRF [15]	$\times$	$\times$	N/A	N/A	35.12	39.21	49.12	31.31	39.77	47.58	$\infty$	4.5	$\infty$	1.0	
BAM [14]	$\times$	$\times$	N/A	N/A	20.43	48.44	53.82	27.24	48.29	52.42	$\infty$	4.5	$\infty$	1.0	
Wavmark [16]	$\checkmark$	$\times$	N/A	N/A	40.22	40.22	50.22	40.00	40.00	50.00	36.9	4.23	$\infty$	1.0	
Audioseal [13]	$\checkmark$	$\times$	N/A	N/A	93.68	98.23	99.02	91.78	97.43	98.69	26.5	4.39	$\infty$	1.0	
Ours	$\checkmark$	$\checkmark$	17.82	3.18	97.02	99.89	99.95	95.40	98.28	98.83	$\infty$	4.5	41.53	0.98	

Table 2: **Impact of Masking Strategy** “No Mask” refers to training without a mask. Our masking strategies outperform “No Mask”, showing their effectiveness.

Masking Strategy	Audio Recovery		Tamper Localization		Visual Quality	
	SNR $\uparrow$	PESQ $\uparrow$	IoU $\uparrow$	AUC $\uparrow$	PSNR $\uparrow$	SSIM $\uparrow$
No Mask	4.63	1.53	60.01	86.10	<b>42.56</b>	<b>0.99</b>
Facial Mask	<b>18.17</b>	<b>2.73</b>	<b>95.19</b>	<b>99.26</b>	41.53	0.98
Random Mask	17.41	2.36	92.29	98.88	40.64	0.98

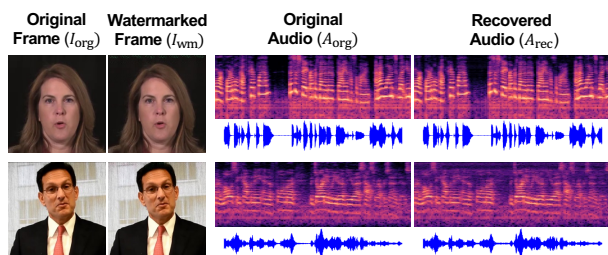


Figure 2: **Qualitative Examples** The watermarked frames and recovered audio closely resemble the original AVS, ensuring imperceptible embedding and authentic audio recovery.

performs the best SOTA method, Audioseal, in both settings and surpasses other baselines by a large margin. Unlike Wavmark and Audioseal, which degrade audio quality by embedding watermarks directly in the audio stream, our approach preserves audio integrity by embedding the watermark into the visual input while maintaining high visual quality. Fig. 2 shows that this process has minimal impact on visual quality.

**Authentic Audio Recovery** Table 1 shows that, despite their tamper localization capabilities, All baseline models are unable to perform AAR. In contrast, our model, leveraging a robust cross-modal watermarking technique, successfully recovers the authentic audio, achieving an SNR of 17.82 and a PESQ of 3.18 from SAVFs, demonstrating outstanding AAR performances. This is confirmed by the qualitative examples in Fig. 2.

**Effects of Masking Strategies** Table 2 compares different masking strategies during training, highlighting their impact on AAR and TLA. Training without masking fails to recover authentic audio, as watermarks are fragile to lip synchronization forgery, making masking essential for AAR. Both random and facial masking improve robustness, with facial masking achieving higher performances since its masked region aligns with the actual tampered areas, enabling more effective signal embedding. In terms of visual quality, training without masking preserves visual fidelity best, as it allows for compact signal embedding but lacks robustness to attacks. Among masking strategies, facial masking maintains better visual quality than random masking, as it introduces less redundancy due to its more predictable regions, while random masking increases uncertainty, leading to more dispersed watermark embedding.

**Robustness to Various Lip Synchronization Methods** We evaluate our model’s robustness against three lip synchroniza-

Table 3: **Comparison of Different Lip Synchronization Methods.** Lip synchronization is simulated with Wav2Lip [7], Diff2Lip [8], and MuseTalk [31]. The watermark is effectively extracted after lip synchronization.

Lip synchronization Model	Audio Recovery		Tamper Localization		
	SNR $\uparrow$	PESQ $\uparrow$	IoU $\uparrow$	AP $\uparrow$	AUC $\uparrow$
None	28.20	3.49	97.13	99.71	99.80
Wav2Lip [7]	16.06	2.91	92.97	95.23	98.29
Diff2Lip [8]	18.17	2.73	95.19	99.02	99.26
MuseTalk [31]	17.82	3.18	95.40	98.28	98.83

Table 4: **Domain Generalization to Unseen Domains during Training.** Comparisons of our method trained on human-associated [28] and non-human [32, 33] datasets, demonstrating its domain generalization capability.

Training Dataset	Audio Recovery		Tamper Localization		Visual Quality	
	SNR $\uparrow$	PESQ $\uparrow$	IoU $\uparrow$	AUC $\uparrow$	PSNR $\uparrow$	SSIM $\uparrow$
HDTF [28]	17.41	2.36	92.29	98.88	40.64	0.98
V190k + FMA [32, 33]	15.40	1.82	88.26	97.72	40.49	0.98

tion methods [7, 8, 31]. As shown in Table 3, performance is highest without lip synchronization, preserving the embedded watermark. While lip synchronization causes some degradation, speech remains intelligible (refer to Supplementary Material), and scores remain relatively high compared to the failed model in Table 2. Additionally, scores for TLA consistently exceed 90%. These results highlight our model’s effectiveness in SAVF scenarios and demonstrate its reliability under various adversarial conditions.

**Learning from Non-Human Datasets** Training networks with human-associated datasets often raises ethical concerns, particularly regarding privacy. To mitigate these issues, we train our model using non-human datasets that exclude human facial imagery and human voice. Specifically, we use the Vimeo-90k dataset [32], combined with the Free Music Archive (FMA) dataset [33]. As shown in Table 4, our model trained on this dataset achieves slightly lower but comparable performance to the model trained directly on HDTF in both AAR and TLA. These results demonstrate that our approach can be effectively trained on datasets from entirely different domains, addressing privacy concerns without significant performance degradation.

## 5. Conclusion

In this paper, we introduced a novel task of recovering authentic audio from SAVFs, moving beyond mere detection and localization. To achieve this, we propose cross-modal watermarking method not only localizing tampered regions but also recovering authentic audio. Our model demonstrated state-of-the-art localization performance while effectively recovering authentic audio. Notably, our approach remains effective without training on human faces or voices, ensuring privacy compliance. This practical solution combats misinformation and preserves content authenticity, fostering a safer multimedia ecosystem.

## 6. Acknowledgements

This research was supported by IITP grants (IITP-2025-RS-2020-II201819, IITP-2025-RS-2024-00436857, IITP-2025-RS-2024-00398115, IITP-2025-RS-2025-02263754, IITP-2025-RS-2025-02304828), and the KOCCA grant (RS-2024-00345025) funded by the Korea government (MSIT, MOE and MSCT).

## 7. References

- [1] M. Le, A. Vyas, B. Shi, B. Karrer, L. Sari, R. Moritz, M. Williamson, V. Manohar, Y. Adi, J. Mahadeokar *et al.*, “Voicebox: Text-guided multilingual universal speech generation at scale,” *NeurIPS*, vol. 36, 2024.
- [2] J. Kim, J. Kong, and J. Son, “Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech,” in *ICML*. PMLR, 2021, pp. 5530–5540.
- [3] R. Shimizu, R. Yamamoto, M. Kawamura, Y. Shirahata, H. Doi, T. Komatsu, and K. Tachibana, “Prompttts++: Controlling speaker identity in prompt-based text-to-speech using natural language descriptions,” in *ICASSP*. IEEE, 2024, pp. 12 672–12 676.
- [4] Z. Qin, W. Zhao, X. Yu, and X. Sun, “Openvoice: Versatile instant voice cloning,” *arXiv preprint arXiv:2312.01479*, 2023.
- [5] H. Zhang, T. Yuan, J. Chen, X. Li, R. Zheng, Y. Huang, X. Chen, E. Gong, Z. Chen, X. Hu *et al.*, “Paddlespeech: An easy-to-use all-in-one speech toolkit,” in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*. Association for Computational Linguistics, 2022.
- [6] G. Ruggiero, E. Zovato, L. Di Caro, and V. Pollet, “Voice cloning: a multi-speaker text-to-speech synthesis approach based on transfer learning,” *arXiv preprint arXiv:2102.05630*, 2021.
- [7] K. R. Prajwal, R. Mukhopadhyay, V. P. Namboodiri, and C. 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*, ser. MM ’20. New York, NY, USA: Association for Computing Machinery, 2020, p. 484–492. [Online]. Available: <https://doi.org/10.1145/3394171.3413532>
- [8] S. Mukhopadhyay, S. Suri, R. T. Gadde, and A. Shrivastava, “Diff2lip: Audio conditioned diffusion models for lip-synchronization,” in *WACV*, January 2024, pp. 5292–5302.
- [9] H. Zhou, Y. Sun, W. Wu, C. C. Loy, X. Wang, and Z. Liu, “Pose-controllable talking face generation by implicitly modularized audio-visual representation,” in *CVPR*, 2021.
- [10] X. Cheng, R. Huang, L. Li, T. Jin, Z. Wang, A. Yin, M. Li, X. Duan, Z. Zhao *et al.*, “Transface: Unit-based audio-visual speech synthesizer for talking head translation,” *arXiv preprint arXiv:2312.15197*, 2023.
- [11] X. Yang, X. Cheng, D. Fu, M. Fang, J. Zuo, S. Ji, Z. Zhao, and J. Tao, “Synctalklip: Highly synchronized lip-readable speaker generation with multi-task learning,” in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 8149–8158.
- [12] X. Zhang, R. Li, J. Yu, Y. Xu, W. Li, and J. Zhang, “Editguard: Versatile image watermarking for tamper localization and copyright protection,” in *CVPR*, 2024, pp. 11 964–11 974.
- [13] R. San Roman, P. Fernandez, H. Elshahar, A. D’efosse, T. Furon, and T. Tran, “Proactive detection of voice cloning with localized watermarking,” *ICML*, 2024.
- [14] J. Zhong, B. Li, and J. Yi, “Enhancing partially spoofed audio localization with boundary-aware attention mechanism,” *arXiv preprint arXiv:2407.21611*, 2024.
- [15] J. Wu, W. Lu, X. Luo, R. Yang, Q. Wang, and X. Cao, “Coarse-to-fine proposal refinement framework for audio temporal forgery detection and localization,” in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 7395–7403.
- [16] G. Chen, Y. Wu, S. Liu, T. Liu, X. Du, and F. Wei, “Wavmark: Watermarking for audio generation,” 2023.
- [17] C.-K. Chan and L. Cheng, “Hiding data in images by simple lsb substitution,” *Pattern Recognition*, vol. 37, no. 3, pp. 469–474, 2004. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S003132030300284X>
- [18] K. Wong, K. Tanaka, K. Takagi, and Y. Nakajima, “Complete video quality-preserving data hiding,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, no. 10, pp. 1499–1512, 2009.
- [19] K. A. Zhang, A. Cuesta-Infante, and K. Veeramachaneni, “Steganogan: High capacity image steganography with gans,” *arXiv preprint arXiv:1901.03892*, 2019. [Online]. Available: <https://arxiv.org/abs/1901.03892>
- [20] L. Dinh, D. Krueger, and Y. Bengio, “Nice: Non-linear independent components estimation,” 2015. [Online]. Available: <https://arxiv.org/abs/1410.8516>
- [21] J. Jing, X. Deng, M. Xu, J. Wang, and Z. Guan, “Hinnet: Deep image hiding by invertible network,” in *ICCV*, 2021, pp. 4733–4742.
- [22] C. Mou, Y. Xu, J. Song, C. Zhao, B. Ghanem, and J. Zhang, “Large-capacity and flexible video steganography via invertible neural network,” in *CVPR*, 2023, pp. 22 606–22 615.
- [23] L. Zhao, H. Li, X. Ning, and X. Jiang, “Thiming: Cross-modal steganography for presenting talking heads in images,” in *WACV*, 2024, pp. 5553–5562.
- [24] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, “Densely connected convolutional networks,” in *cvpr*, 2017.
- [25] S. Hershey, S. Chaudhuri, D. P. Ellis, J. F. Gemmeke, A. Jansen, R. C. Moore, M. Plakal, D. Platt, R. A. Saurous, B. Seybold *et al.*, “Cnn architectures for large-scale audio classification,” in *2017 IEEE international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2017, pp. 131–135.
- [26] A. van den Oord, Y. Li, and O. Vinyals, “Representation learning with contrastive predictive coding,” 2019. [Online]. Available: <https://arxiv.org/abs/1807.03748>
- [27] S. I. Serengil and A. Ozpinar, “Lightface: A hybrid deep face recognition framework,” in *2020 Innovations in Intelligent Systems and Applications Conference (ASYU)*. IEEE, 2020, pp. 23–27. [Online]. Available: <https://ieeexplore.ieee.org/document/9259802>
- [28] Z. Zhang, L. Li, Y. Ding, and C. Fan, “Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset,” in *CVPR*, 2021, pp. 3661–3670.
- [29] T. Afouras, J. S. Chung, and A. Zisserman, “Lrs3-ted: a large-scale dataset for visual speech recognition,” *arXiv preprint arXiv:1809.00496*, 2018.
- [30] X. Zhang, Y. Xu, R. Li, J. Yu, W. Li, Z. Xu, and J. Zhang, “V2a-mark: Versatile deep visual-audio watermarking for manipulation localization and copyright protection,” in *Proceedings of the 32nd ACM International Conference on Multimedia*, 2024, pp. 9818–9827.
- [31] Y. Zhang, M. Liu, Z. Chen, B. Wu, Y. Zeng, C. Zhan, Y. He, J. Huang, and W. Zhou, “Musetalk: Real-time high quality lip synchronization with latent space inpainting,” *arxiv*, 2024.
- [32] T. Xue, B. Chen, J. Wu, D. Wei, and W. T. Freeman, “Video enhancement with task-oriented flow,” *International Journal of Computer Vision*, vol. 127, pp. 1106–1125, 2019.
- [33] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, “FMA: A dataset for music analysis,” in *18th International Society for Music Information Retrieval Conference (ISMIR)*, 2017. [Online]. Available: <https://arxiv.org/abs/1612.01840>