DEFENDING FACES WITHOUT TOUCHING PIXELS: AUDIO-ONLY PROTECTION AGAINST PERSONALIZED TALKING FACE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

The rapid development of talking face generation (TFG) has raised growing concerns about identity misuse. In particular, audio-driven 3D TFG can reconstruct a reusable personalized 3D portrait from a reference video and animate it with arbitrary audio, making identity impersonation increasingly accessible. To counter such unauthorized reuse, existing proactive defenses predominantly inject protective perturbations into visual content to disrupt generative identity acquisition. However, these defenses face a sharp tradeoff between defense effectiveness and perceptual quality. Perturbing facial regions can noticeably degrade portrait appearance due to the high perceptual sensitivity of human faces, whereas shifting perturbations toward less visually sensitive regions or the background makes them more vulnerable to resizing, compression, and background replacement. To mitigate these limitations, we propose **VoxGuard**, a perceptually constrained audio-only defense that disrupts the learning of correspondences between audio and identity-specific facial motion during personalization. Instead of perturbing visual content, VoxGuard modifies only the audio track of the reference video, leaving visual frames unchanged and thereby avoiding modification of visually sensitive facial regions. To further reduce perceptual distortion, we employ psychoacoustic masking to guide perturbations toward perceptually masked regions. We further optimize across heterogeneous audio encoders to enhance black-box transferability and introduce continual optimization to mitigate interference among encoder objectives while preserving previously established representation shifts. Extensive experiments demonstrate that VoxGuard effectively disrupts unauthorized 3D TFG while limiting perceptual degradation, exhibits black-box transferability, and remains effective after common audio transformations, including MP3 and AAC compression and audio resampling.

1 INTRODUCTION

The rapid development of talking face generation (TFG) has enabled increasingly realistic facial animation from audio and portrait media. In particular, audio-driven 3D TFG can reconstruct a reusable personalized 3D portrait from a portrait video and animate it with arbitrary audio (Li et al., 2023; Peng et al., 2024; Li et al., 2025a). Recent methods further reduce the required reference footage to a few seconds (Li et al., 2025a), lowering the amount of personal media needed for portrait reconstruction. Although these capabilities support digital humans and virtual communication, they also make identity impersonation increasingly accessible. A publicly shared video can become the basis for a reusable digital portrait that depicts a person speaking words they never said. Protecting portrait media before unauthorized collection and reuse is therefore an important challenge.

Proactive defenses address this challenge by embedding protective perturbations into the portrait videos before they are shared. Existing approaches predominantly operate in the visual domain, disrupting the extraction and reuse of visual identity information (Salman et al., 2023; Wang et al., 2025; Jeong et al., 2025b). Recent work extends visual protection to the acquisition of personalized 3D portraits (Sun et al., 2025). Other methods perturb reference portraits to suppress audio control in diffusion models for TFG (Gan et al., 2025). Despite targeting different components,

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.

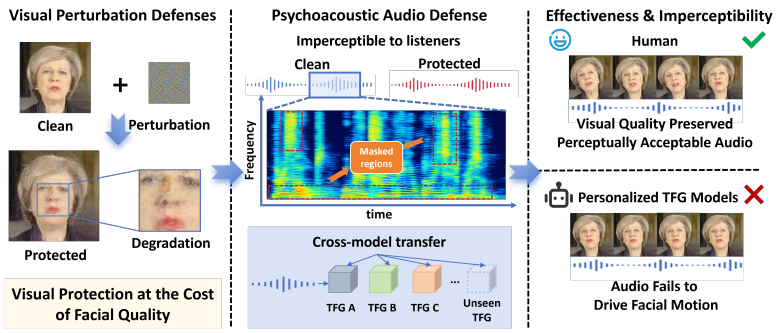


Figure 1: Motivation and overview of our audio track proactive defense. Instead of perturbing facial pixels, our method modifies only the reference audio to preserve visual quality while disrupting the learning of audio driven facial motion. Psychoacoustic masking and heterogeneous audio representations further improve perceptual quality and black-box transferability.

these approaches share a common intervention: modifying the visual content used by the generation system.

However, visual defenses face a sharp tradeoff between defense effectiveness and perceptual quality. Perturbing facial regions can noticeably degrade portrait appearance, as human face perception is highly sensitive to subtle changes in facial features and their spatial configuration (Tanaka & Sengco, 1997; Namdar et al., 2015). Shifting perturbations toward less visually sensitive regions or the background can alleviate this perceptual cost, but introduces additional vulnerabilities. For example, background replacement can directly remove perturbations outside the face, while common media processing operations such as resizing and compression can attenuate protective perturbations before the content reaches a generation system (Qu et al., 2024). Although recent methods improve robustness to such transformations (Jeong et al., 2025b; Sun et al., 2025), recent evidence shows that protective perturbations can still be weakened or invalidated during media distribution in practice (Sun et al., 2026). These limitations motivate an alternative form of protection that avoids modifying visually sensitive facial regions altogether. As illustrated in Fig. 1, we therefore shift the intervention point from visually sensitive facial content to the audio track.

To this end, we propose **VoxGuard**, a perceptually constrained audio-only defense that targets correspondence learning during the personalization of audio-driven 3D TFG. Audio-driven 3D TFG learns correspondences between audio and identity-specific facial motion from the reference video (Li et al., 2023; Peng et al., 2024; Li et al., 2025a). VoxGuard perturbs only the reference audio to disrupt the reliable acquisition of these correspondences, while leaving visual frames unchanged and thereby avoiding modification of visually sensitive facial regions. Specifically, VoxGuard integrates three complementary strategies. First, psychoacoustic masking adaptively constrains perturbation energy according to local auditory masking thresholds, reducing audible distortion while preserving greater optimization freedom in perceptually masked time-frequency regions (Qin et al., 2019; Schönherr et al., 2018). Second, perturbations are optimized across heterogeneous audio encoders to reduce dependence on any single representation space, thereby improving broader generalization, including black-box transfer to unseen audio representations. Third, continual optimization is introduced to mitigate interference among encoder objectives. Rather than optimizing all objectives jointly from the outset, encoder objectives are introduced progressively while previously established representation shifts are explicitly preserved, reducing forgetting caused by competing optimization goals (Yu et al., 2020).

Our main contributions are summarized as follows:

- We recast proactive protection against personalized audio-driven 3D TFG by targeting audio-to-motion correspondence learning rather than visual identity acquisition, enabling defense without modifying visual frames.
- We propose **VoxGuard**, an audio-only defense that combines psychoacoustic masking, heterogeneous encoder optimization, and continual optimization for perceptually constrained and transferable protection under black-box settings.

- Extensive evaluations demonstrate effective transfer across identities and audio representations, robustness to common audio transformations, and higher overall perceptual acceptability than the evaluated visual defenses.

2 Related Work

2.1 Audio-driven Talking Face Generation

Audio-driven talking face generation (TFG) aims to synthesize realistic talking portraits by animating facial motions according to audio signals. Early methods mainly relied on 2D image generation or warping-based pipelines (Chen et al., 2019; 2018; Prajwal et al., 2020), which achieved plausible lip synchronization but often struggled to model personalized head movement, fine-grained facial dynamics, and multi-view consistency. Recent advances in 3D representations have substantially improved the realism and controllability of TFG. In particular, Neural Radiance Fields (NeRF) (Athar et al., 2022; Bi et al., 2024; Guo et al., 2021; Tang et al., 2025; Li et al., 2023; Ye et al., 2023; Peng et al., 2024) and 3D Gaussian Splatting (3DGS) (Cho et al., 2024; Li et al., 2024; 2025a) enable the reconstruction of subject-specific 3D portraits from monocular reference videos, leading to stronger identity consistency, more accurate lip synchronization, and better view-consistent rendering. Compared with generalized one-shot portrait animation methods (Li et al., 2025b; Ye et al., 2024), these personalized 3D-field approaches are particularly powerful because they can recover reusable identity-aware 3D representations and support highly realistic audio-driven reenactment However, once a target person's reference video is collected, personalized 3D TFG systems can reconstruct a reusable digital portrait that can be animated with arbitrary audio, creating substantial risks of identity misuse. We therefore focus on proactive defense against audio-driven 3D TFG systems.

2.2 Talking Face Defense

Recent studies have explored proactive defenses against unauthorized portrait manipulation including TFG, with most methods operating in the visual domain by perturbing images or videos to disrupt identity extraction, reconstruction, editing, or generation (Liang et al., 2023; Liang & Wu, 2023; Salman et al., 2023; Jeong et al., 2025a; Wang et al., 2025; Sun et al., 2025; Gan et al., 2025). Although effective, visual perturbations directly modify facial appearance and can be weakened by common media transformations such as resizing and compression (Qu et al., 2024; Sun et al., 2026). Our work instead revisits where protection should be imposed in audio driven 3D TFG. During personalization, the model learns correspondences between audio and identity specific facial motion from the reference video. The audio track of the reference video is therefore perturbed to disrupt this correspondence learning, while the visual frames are left unchanged. The key distinction from prior work is thus defined not merely by the modality being perturbed, but by the information acquisition process being targeted.

3 Preliminaries

This section introduces the underlying mechanism of personalized 3D TFG models targeted by our defense, where facial motion is learned and driven from paired audio and visual observations (Li et al., 2023; Peng et al., 2024; Li et al., 2025a). Psychoacoustic masking is further introduced as the perceptual basis for reducing the audible impact of defensive perturbations applied to the audio track (Qin et al., 2019; Schönherr et al., 2018).

3.1 Audio Driven Talking Face Generation in 3D Fields

Audio driven 3D TFG reconstructs a personalized 3D representation of a target face from a monocular reference video and animates it according to audio. Compared with 2D image based generation, the explicit or implicit 3D representation supports more consistent facial geometry and rendering across viewpoints. The reference video provides both visual observations for reconstructing the target identity and paired audio for learning how audio corresponds to facial motion. During personalization, visual observations provide identity appearance, facial geometry, expression, and head

pose, while the paired audio provides the acoustic conditions associated with the observed facial motions. The model therefore learns not only a reusable 3D portrait, but also the correspondence between audio representations and identity specific facial motion.

Although existing methods adopt different 3D representations, they share this general conditional rendering principle. NeRF based methods represent the portrait as a continuous radiance field, where density and color are queried at sampled 3D locations and conditioned on audio features before volume rendering. In contrast, 3DGS based methods represent the portrait using explicit Gaussian primitives and model speaking motion through audio conditioned deformation of these primitives. Despite these architectural differences, both paradigms ultimately learn a mapping from audio representations to identity specific facial motion. This process can be abstracted as

$$\mathbf{a}_t = \mathcal{E}(\mathbf{x}_t), \qquad \mathcal{S}_t = \Phi(\mathcal{S}_{id}, \mathbf{a}_t, \mathbf{p}_t, \mathbf{e}_t), \qquad \hat{\mathbf{I}}_t = \mathcal{R}(\mathcal{S}_t), \tag{1}$$

where $\mathbf{x}_t$ denotes the audio input, $\mathcal{E}$ is the audio encoder, $\mathcal{S}_{id}$ denotes the personalized 3D representation, and $\mathbf{p}_t$ and $\mathbf{e}_t$ denote pose and expression related conditions. The function Φ maps these conditions to a dynamic 3D state $\mathcal{S}_t$, which is rendered by $\mathcal{R}$ to produce the output frame $\hat{\mathbf{I}}_t$. The learned dependence between $\mathbf{a}_t$ and the facial dynamics encoded in $\mathcal{S}_t$ constitutes the correspondence targeted by our defense.

3.2 Psychoacoustic Masking

Given that human auditory sensitivity is nonuniform across time and frequency, strong spectral components can raise local hearing thresholds in neighboring regions, making weaker signals less perceptible Psychoacoustic masking is therefore employed to reduce the perceptual impact of defensive perturbations applied to the audio track (Qin et al., 2019; Schönherr et al., 2018). Given a source audio signal $\mathbf{x}$, the global masking threshold is computed from its time and frequency structure as

$$\Theta_{\mathbf{x}}(t, f) = 10 \log_{10} \left(10^{T_q(f)/10} + \sum_{i \in \mathcal{M}_t} 10^{T_i(t,f)/10} \right), \tag{2}$$

where $T_q(f)$ denotes the absolute threshold of hearing, $\mathcal{M}_t$ is the set of detected maskers in time frame t, and $T_i(t, f)$ denotes the local masking threshold induced by the ith masker. Perturbations below this threshold are expected to be less perceptible to human listeners. The resulting masking threshold is used to construct the perceptual constraint in Sec. 4.1. Further details on its computation and implementation are provided in Appendix A.

3.3 Threat Model and Defense Objective

We consider a user who shares a portrait video $\mathcal{V} = \{\mathbf{I}_{1:T}, \mathbf{x}\}$ on public platforms, where $\mathbf{I}_{1:T}$ denotes the visual frames and $\mathbf{x}$ denotes the corresponding audio track. An unauthorized adversary may collect this video and use it to personalize an audio driven 3D TFG model, which can subsequently animate a reusable portrait of the target identity with arbitrary audio. Our goal is to protect the reference video by perturbing only its audio track while leaving all visual frames unchanged. Specifically, a defensive perturbation $\boldsymbol{\delta}$ is added as

$$\widetilde{\mathbf{x}} = \mathrm{clip}\left(\mathbf{x} + \boldsymbol{\delta}, -1, 1\right), \tag{3}$$

yielding the protected reference video $\widetilde{\mathcal{V}} = \{\mathbf{I}_{1:T}, \widetilde{\mathbf{x}}\}$. The perturbation is optimized to disrupt the reliable learning of correspondences between audio and identity specific facial motion during personalization, while reducing audible distortion.

4 Method

Our goal is to disrupt the learning of correspondences between audio and identity specific facial motion during the personalization of 3D TFG models, while leaving the visual frames unchanged. Although adding sufficiently strong Gaussian noise directly to the audio track can disrupt the correspondence between audio and facial motion, it inevitably degrades the perceptual quality of the protected audio. Reducing the magnitude of the noise alleviates perceptual distortion but also weakens its ability to disrupt correspondence learning. To establish a controllable defense objective,

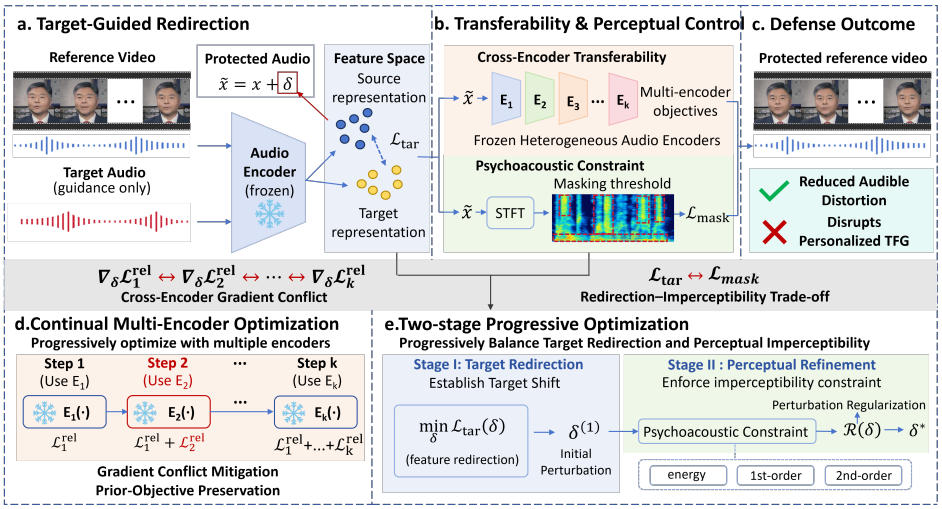


Figure 2: Overview of our proactive audio defense.

target guided feature redirection is introduced under an explicit perturbation budget, as illustrated in Fig 2(a). Given the source audio x, we introduce an unrelated target audio x^{tar} to define the representation redirection objective. For consistency across experiments, x^{tar} is a fixed, duration-matched utterance spoken by a speaker different from the protected identity and containing different linguistic content from x. The same target is retained throughout optimization without resampling. The protected audio is defined as

$$\widetilde{\mathbf{x}} = \text{clip}\left(\mathbf{x} + \boldsymbol{\delta}, -1, 1\right), \qquad \|\boldsymbol{\delta}\|_\infty \leq \epsilon, \tag{4}$$

where $\boldsymbol{\delta}$ denotes the defensive perturbation and ϵ bounds its maximum sample wise amplitude. For a frozen audio encoder $\mathcal{E}$, let $\mathbf{f}_t^{adv}$ and $\mathbf{f}_t^{tar}$ denote the encoded features of the protected and target audio, respectively, at temporal position t. The target redirection objective is defined as

$$\mathcal{L}_{tar} = \frac{1}{T}\sum_{t=1}^{T} \left\|\text{Norm}\left(\mathbf{f}_t^{adv}\right) - \text{Norm}\left(\mathbf{f}_t^{tar}\right)\right\|_2^2. \tag{5}$$

By minimizing $\mathcal{L}_{tar}$ under the perturbation constraint, the acoustic representation of the protected audio is redirected toward that of the unrelated target audio while preventing unconstrained waveform modification. This representation shift interferes with the correspondence between the observed facial motion and its associated audio during personalization. The perturbation is further constrained according to human auditory perception in Sec. 4.1.

However, effective feature redirection alone does not ensure a practical defense. The resulting perturbation should incur limited perceptual distortion, transfer to unseen audio representations under black-box settings, and preserve previously established representation shifts when multiple objectives are optimized together. As illustrated in Fig. 2, these three challenges are addressed by psychoacoustic masking, heterogeneous encoder optimization, and conflict aware progressive optimization, respectively.

4.1 PSYCHOACOUSTIC MASKING FOR PERCEPTUAL CONTROL

Although the perturbation budget limits the overall magnitude of the defensive perturbation, a uniform amplitude constraint does not account for the nonuniform sensitivity of human hearing across time and frequency. Perturbation energy may therefore remain perceptible when concentrated in sensitive regions. To further reduce its perceptual impact, the psychoacoustic masking mechanism introduced in Sec. 3.2 is incorporated to adaptively constrain the perturbation according to the local auditory sensitivity of the source audio, as illustrated in Fig. 2(b).

Given the masking threshold $\Theta_{\mathbf{x}}(t, f)$ computed from the source audio $\mathbf{x}$, the power spectral density of the perturbation is normalized using the clean audio within the same time frame as

$$\bar{P}_\delta(t, f) = P_\delta(t, f) - \max_{f'} P_{\mathbf{x}}(t, f') + P_{\text{top}}, \tag{6}$$

where $P_\delta(t, f)$ and $P_{\mathbf{x}}(t, f)$ denote the perturbation and clean audio power spectral densities in dB, respectively, and P_{top} denotes the reference sound pressure level. The same clean audio reference and P_{top} are used in computing $\Theta_{\mathbf{x}}(t, f)$, placing $\bar{P}_\delta(t, f)$ and $\Theta_{\mathbf{x}}(t, f)$ on the same normalized dB scale and allowing them to be directly compared. The psychoacoustic masking loss is then defined as

$$\mathcal{L}_{\text{mask}} = \frac{1}{TF} \sum_{t=1}^{T} \sum_{f=1}^{F} \left[\bar{P}_\delta(t, f) - (\Theta_{\mathbf{x}}(t, f) - m) \right]_+, \tag{7}$$

where T and F denote the numbers of time frames and frequency bins, respectively, $[u]_+ = \max(u, 0)$, and m denotes a perceptual safety margin. The margin m places the effective constraint below the estimated masking threshold to provide a more conservative perceptual bound.

During the optimization process, only perturbation components exceeding this local masking boundary are penalized. Consequently, greater perturbation energy is permitted in strongly masked regions, while tighter constraints are imposed in perceptually sensitive regions. Compared with uniformly reducing the perturbation magnitude, this perceptually adaptive constraint preserves greater optimization freedom where perturbations are less audible, thereby reducing perceptual distortion without unnecessarily weakening target redirection.

4.2 Heterogeneous Encoder Optimization for Transfer to Unseen Encoders

Target guided redirection optimized against a single audio encoder can become dependent on its specific representation space. Under black-box settings, however, the audio encoder employed by the target TFG model is generally unknown. A perturbation overfitted to one encoder may therefore exhibit substantially reduced effectiveness when transferred to unseen audio representations. To improve black-box transferability, a set of frozen heterogeneous audio encoders $\{\mathcal{E}_i\}_{i=1}^{K}$ is introduced, as illustrated in Fig. 2(b).

For each encoder $\mathcal{E}_i$, the target redirection loss $\mathcal{L}_i$ is computed between the protected audio $\widetilde{\mathbf{x}}$ and the target audio $\mathbf{x}^{tar}$. Since heterogeneous encoders can differ substantially in feature dimension, representation distribution, and raw distance scale, their losses cannot be directly compared. The initial distance between the source audio and the target audio is therefore used as an encoder specific reference, yielding the relative redirection loss

$$\mathcal{L}_i^{\text{rel}} = \frac{\mathcal{L}_i\left(\widetilde{\mathbf{x}}, \mathbf{x}^{tar}\right)}{\mathcal{L}_i\left(\mathbf{x}, \mathbf{x}^{tar}\right)}. \tag{8}$$

This normalization expresses the redirection progress of each encoder relative to its own initial representation distance, reducing the influence of incompatible raw loss scales across encoders. Target redirection can therefore be enforced across multiple heterogeneous representation spaces rather than a single encoder specific space. Dependence on any particular audio representation is consequently reduced, which provides the basis for transfer to unseen audio encoders and generation pipelines.

4.3 Continual Optimization

Continual Multi Encoder Optimization. Jointly optimizing heterogeneous encoder objectives can lead to gradient interference, causing newly introduced objectives to degrade representation shifts established on earlier encoders. We therefore introduce encoder objectives progressively while explicitly limiting forgetting. For block k, let a_k denote the newly introduced encoder, $\mathcal{P}_k$ the previously optimized encoders, and $\mathcal{A}_k = \mathcal{P}_k \cup \{a_k\}$ the active encoder set. The active encoder objectives are adaptively weighted as

$$L_k^{\text{enc}} = \sum_{i \in \mathcal{A}_k} w_{i,k} L_i^{\text{rel}}, \tag{9}$$

where $w_{i,k}$ balances optimization progress across active encoders. The weighted objectives are used to construct an RMS balanced update direction, while previously optimized encoders are protected by

$$L_j^{\mathrm{rel}}(\tilde{x}) \le c_{j,k}, \qquad \forall j \in \mathcal{P}_k, \qquad \|\delta\|_\infty \le \epsilon. \tag{10}$$

Conflicting gradient components are projected against the protected objectives, followed by retention feasible backtracking. Further optimization details are provided in Appendix B.

Two Stage Defense Preserving Refinement. Introducing perceptual constraints from the beginning can restrict the optimization space before stable representation shifts are established. We therefore adopt a two stage strategy. The first stage performs continual multi encoder optimization to establish transferable representation shifts. Starting from the resulting perturbation $\delta^{(1)}$, the second stage introduces psychoacoustic and temporal constraints while retaining the encoder redirection update.

Let g_k^{enc} denote the RMS balanced update direction constructed from the active encoder objectives. The Stage 2 update direction before retention projection is

$$g_k^{(2)} = g_k^{\mathrm{enc}} + \nabla_\theta \left[\lambda_{\mathrm{mask}} L_{\mathrm{mask}} + R(\delta)\right], \tag{11}$$

where $\delta = \epsilon \tanh(\theta)$ and $R(\delta)$ constrains perturbation energy and temporal variation. Retaining g_k^{enc} during perceptual refinement prevents the previously established representation shifts from being erased. The same conflict projection and retention feasible backtracking are applied to the combined update direction. Further optimization details are provided in Appendix B.

5 Experiments

Experimental Setup. We evaluate two representative personalized 3D TFG paradigms, the NeRF-based SyncTalk (Peng et al., 2024) and the 3DGS-based InsTaG (Li et al., 2025a), using AVE, Esperanto, HuBERT, and DeepSpeech as audio encoders. AVE, Esperanto, and HuBERT serve as surrogates, while DeepSpeech is fully held out for black-box evaluation. Talk9 (Li et al., 2025b), a benchmark widely used for evaluating TFG quality, is adopted as the primary test set, with 40 HDTF (Zhang et al., 2021) identities for identity generalization. We compare against AdvDM (Liang et al., 2023), PhotoGuard (Salman et al., 2023), SDS (Xue et al., 2023), and Silencer (Gan et al., 2025). Defense effectiveness is measured by Sync-C/D (Chung & Zisserman, 2016) and PSNR/SSIM/FID, where lower Sync-C/PSNR/SSIM and higher Sync-D/FID indicate stronger disruption. Full experimental details are provided in Appendix C.

5.1 Defense Effectiveness, Generalization, and Robustness

Quantitative results. Table 1 evaluates both defense effectiveness and the visual quality of protected reference videos. As shown in Table 1(a), Gaussian noise has little effect across the evaluated personalized 3D TFG settings, whereas our method consistently degrades audio visual synchronization and generated visual quality. This provides evidence that targeted representation redirection interferes with correspondence learning. Compared with the recent visual defense Silencer (Gan et al., 2025) under its original Hallo setting (Xu et al., 2024), our method achieves comparable or stronger disruption while leaving the visual frames unchanged. Table 1(b) highlights the advantage of shifting protection from facial pixels to the audio track. Visual defenses degrade the protected portrait, whereas our method leaves all visual frames unchanged and therefore preserves their visual quality exactly. This difference also partly accounts for the stronger visual degradation observed for Silencer in Table 1(a), with further analysis are in the Appendix.

Qualitative Results. As shown in Fig. 3, models personalized with clean reference videos preserve more consistent variations in mouth opening and lip configuration. In contrast, models personalized with protected reference videos exhibit weakened or abnormal mouth dynamics at the corresponding timestamps. In several frames, the degradation further appears as incomplete rendering in the mouth region, indicating that the learned correspondence between audio representations and identity-specific facial motion has been disrupted during personalization.

Black-Box Encoder Transfer. DeepSpeech is completely held out from perturbation optimization and used only for evaluation, while all personalization and inference settings remain unchanged. As

(a) Defense Effectiveness

Model	Encoder	Method	Sync-C ↓	Sync-D ↑	PSNR ↓	FID ↑
Hallo	Wav2Vec2.0	Silencer-I	3.315	11.131	18.068	205.407
InsTaG	AVE	Clean	6.249	7.841	31.239	6.550
		Gaussian Noise	5.971	8.190	31.201	7.608
		VoxGuard (ours)	1.852	11.233	30.969	10.416
	Esperanto	Clean	4.534	9.377	31.128	7.435
		Gaussian Noise	4.420	9.350	31.032	9.239
		VoxGuard (ours)	3.263	10.341	31.037	7.323
	HuBERT	Clean	4.621	9.164	31.147	6.492
		Gaussian Noise	4.506	9.215	31.128	9.461
		VoxGuard (ours)	2.157	11.070	30.901	7.493
SyncTalk	AVE	Clean	7.329	9.046	30.824	6.335
		Gaussian Noise	6.483	8.020	30.123	6.477
		VoxGuard (ours)	1.861	11.840	29.758	7.079

(b) Protected Visual Quality

Method	FID ↓	SSIM ↑	PSNR ↑
AdvDM(-)	63.40	0.4793	27.48
PhotoGuard	107.53	0.5040	26.54
SDS(+)	127.02	0.5453	27.57
SDS(-)	67.31	0.5477	28.55
SDST(-)	100.80	0.5530	27.55
Silencer-I	78.29	0.5311	28.48
VoxGuard (ours)	0.00	1.0000	∞

Table 1: Quantitative evaluation of defense effectiveness and protected visual quality

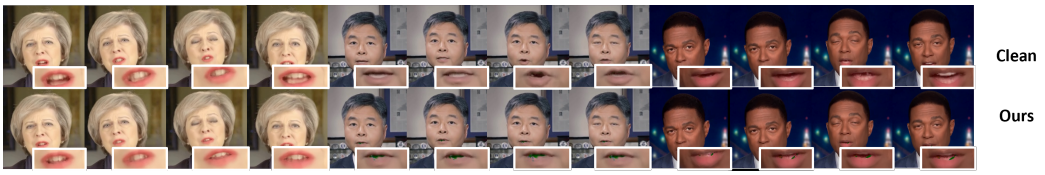


Figure 3: Qualitative comparison on InsTaG with AVE as the audio encoder.

shown in Table 2(a), our defense remains effective under the black-box setting, substantially degrading audio-visual synchronization on both InsTaG and SyncTalk. Despite the unseen audio encoder, its disruption is comparable to Silencer under its original setting, demonstrating strong black-box transferability. Beyond transfer across encoders, we further evaluate generalization across generation pipelines by fully fine tuning MuseTalk 1.5 (Zhang et al., 2024) for each Talk9 identity using clean or protected videos. Fine tuning on protected videos yields weaker audio visual synchronization, providing further evidence that our defense disrupts the learning of correspondences between speech and identity specific facial motion during personalization (Appendix D).

Identity Generalization. We further evaluate identity generalization on 40 randomly sampled identities from HDTF. As shown in Table 2(b), our defense consistently degrades audio-visual synchronization across AVE, Esperanto, and HuBERT, demonstrating robust generalization across identities beyond the primary evaluation set.

Robustness to Common Audio Transformations. We further evaluate robustness to common audio processing operations, including MP3 and AAC compression at 128 kbps and resampling from 22.05 kHz to 16 kHz. As shown in Table 2(c), Sync-C and -D remain close to the untransformed results, indicating that the disruption is largely preserved under common processing.

5.2 Ablation Study

Defense Effectiveness. We first conducted an ablation study focusing solely on the optimization of the AVE encoder. As shown in Table 3(a), removing the retention constrain sharply weakens synchronization disruption, with Sync-C increasing from 1.852 to 5.388 and Sync-D decreasing from 11.233 to 8.423. This confirms that previously established representation shifts must be explicitly preserved during sequential optimization. Removing psychoacoustic masking has little effect on synchronization, consistent with their primary roles in perceptual control. The lower CDPAM (Manocha et al., 2021) without preservation is accompanied by substantially weaker defense, exposing the tradeoff between perceptual quality and disruption strength.

Generalization to Unseen Encoder. DeepSpeech is used only for evaluation. As shown in Table 3(b), multiple encoders improve transfer over a single encoder, while the VoxGuard further outperforms naive joint optimization and optimization without continual preservation. This shows

(a) Unseen DeepSpeech Encoder)

	Sync-C ↓		Sync-D ↑	
Model	**Clean**	**Ours**	**Clean**	**Ours**
InsTaG	4.273	**3.196**	9.597	**10.587**
SyncTalk	4.955	**4.067**	9.362	**10.189**

(b) Identity Generalization on HDTF with InsTaG

	Sync-C ↓		Sync-D ↑	
Encoder	**Clean**	**Ours**	**Clean**	**Ours**
AVE	6.982	**2.793**	7.634	**10.641**
Esperanto	5.423	**4.303**	8.951	**9.666**
HuBERT	5.536	**3.692**	8.811	**10.170**

(c) Audio Transformations on InsTaG with AVE

Metric	**None**	**MP3**	**AAC**	**Resamp.**
Sync-C ↓	1.852	1.944	1.969	2.168
Sync-D ↑	11.233	11.236	11.254	11.065

Table 2: Generalization across encoders and identities, and robustness to audio transformations.

(a) Defense Ablation (InsTaG + AVE)

Metric	w/o Retention	w/o Masking	Full
Sync-C ↓	5.3883	1.9070	1.8520
Sync-D ↑	8.4232	11.2074	11.2330
CDPAM ↓	0.1478	0.3221	0.3185

(b) Generalization Ablation (InsTaG + DeepSpeech)

Metric	Single	Joint	w/o Continual Opt.	VoxGuard
Sync-C ↓	3.9733	3.4555	3.3828	3.1963
Sync-D ↑	9.9007	10.3539	10.3341	10.5874
CDPAM ↓	0.3185	0.2613	0.2686	0.2842

(c) Human Perceptual Evaluation

Method	Overall ↑	Visual ↑	Audio ↑
Clean	4.61	4.65	4.54
VoxGuard (ours)	**3.63**	**4.74**	2.98
Silencer	3.30	3.05	4.62
AdvDM(-)	3.23	2.94	4.71
SDS(-)	3.03	2.72	4.71
SDST(-)	2.46	2.19	4.66
PhotoGuard	2.38	2.04	4.63
Mist	2.33	2.20	4.68
SDS(+)	2.24	2.00	4.60
AdvDM(+)	2.18	1.97	4.64

Table 3: Ablation and human perceptual evaluation.

that representation diversity alone is insufficient and preserving progress across encoder objectives is critical for transfer to unseen representations.

We further analyze the sensitivity to the perturbation budget ϵ and the number of optimization steps in Appendix F, where $\epsilon = 0.02$ and 80 steps per encoder block in each stage provide a balanced tradeoff between defense effectiveness and perceptual quality.

5.3 Perceptual Quality Evaluation

We conduct a subjective study on the perceptual quality of protected reference videos, using overall acceptability as the primary metric for cross-modality comparison and visual/audio quality to characterize modality-specific degradation. As shown in Table 3(c), visual defenses noticeably degrade facial appearance, whereas our audio-only defense leaves the visual content unchanged and shifts the perceptual cost to audio. By preserving facial appearance while constraining audible distortion through psychoacoustic masking, our method achieves higher overall acceptability than the evaluated visual defenses. This result supports our central motivation that directly perturbing visually sensitive facial content can incur substantial perceptual cost. Additional study details are provided in the Appendix E.

6 Conclusion

We revisit proactive protection against personalized audio-driven 3D TFG from a different intervention point. Rather than defending visual identity by perturbing facial pixels, we target the correspondence between speech and identity-specific facial motion that must be learned during personalization. This shift enables an audio-only defense that leaves the visual frames unchanged while still disrupting the acquisition of personalized facial motion. By combining psychoacoustic masking, optimization across heterogeneous audio encoders, and continual optimization, our method further improves perceptual quality while maintaining transferable protection, black-box transferability, and robustness to common audio transformations. Human evaluation also shows that preserving facial appearance leads to higher overall perceptual acceptability than the evaluated visual defenses. These results suggest that effective portrait protection does not necessarily have to come at the cost of portrait appearance, and highlight correspondence learning as a promising intervention point for proactive defense against personalized generative models.

REFERENCES

Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep speech 2: End-to-end speech recognition in english and mandarin. In International conference on machine learning, pp. 173–182. PMLR, 2016.

ShahRukh Athar, Zexiang Xu, Kalyan Sunkavalli, Eli Shechtman, and Zhixin Shu. Rignerf: Fully controllable neural 3d portraits. In Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition, pp. 20364–20373, 2022.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems, 33:12449–12460, 2020.

Benjamin Belmudez and Sebastian Möller. Audiovisual quality integration for interactive communications. EURASIP Journal on Audio, Speech, and Music Processing, 2013(1):24, 2013.

Chongke Bi, Xiaoxing Liu, and Zhilei Liu. Nerf-ad: Neural radiance field with attention-based disentanglement for talking face synthesis. In ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 3490–3494. IEEE, 2024.

Lele Chen, Zhiheng Li, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Lip movements generation at a glance. In Proceedings of the European conference on computer vision (ECCV), pp. 520–535, 2018.

Lele Chen, Ross K Maddox, Zhiyao Duan, and Chenliang Xu. Hierarchical cross-modal talking face generation with dynamic pixel-wise loss. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 7832–7841, 2019.

Kyusun Cho, Joungbin Lee, Heeji Yoon, Yeobin Hong, Jaehoon Ko, Sangjun Ahn, and Seungryong Kim. Gaussiantalker: Real-time talking head synthesis with 3d gaussian splatting. In Proceedings of the 32nd ACM International Conference on Multimedia, pp. 10985–10994, 2024.

Joon Son Chung and Andrew Zisserman. Out of time: automated lip sync in the wild. In Asian conference on computer vision, pp. 251–263. Springer, 2016.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. Unsupervised cross-lingual representation learning for speech recognition. arXiv preprint arXiv:2006.13979, 2020.

Hugo Fastl and Eberhard Zwicker. Psychoacoustics: Facts and models, volume 22. Springer Science & Business Media, 2006.

Yuan Gan, Jiaxu Miao, Yunze Wang, and Yi Yang. Silence is golden: Leveraging adversarial examples to nullify audio control in ldm-based talking-head generation. In Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 13434–13444, 2025.

Yudong Guo, Keyu Chen, Sen Liang, Yong-Jin Liu, Hujun Bao, and Juyong Zhang. Ad-nerf: Audio driven neural radiance fields for talking head synthesis. In Proceedings of the IEEE/CVF international conference on computer vision, pp. 5784–5794, 2021.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems, 30, 2017.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. HuBERT: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 29:3451–3460, 2021. doi: 10.1109/TASLP.2021.3122291.

Jaehwan Jeong, Sumin In, Sieun Kim, Hannie Shin, Jongheon Jeong, Sang Ho Yoon, Jaewook Chung, and Sangpil Kim. Faceshield: Defending facial image against deepfake threats. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10364–10374, 2025a.

Jaehwan Jeong, Sumin In, Sieun Kim, Hannie Shin, Jongheon Jeong, Sang Ho Yoon, Jaewook Chung, and Sangpil Kim. Faceshield: Defending facial image against deepfake threats. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10364–10374, 2025b.

Jiahe Li, Jiawei Zhang, Xiao Bai, Jun Zhou, and Lin Gu. Efficient region-aware neural radiance fields for high-fidelity talking portrait synthesis. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7568–7578, 2023.

Jiahe Li, Jiawei Zhang, Xiao Bai, Jin Zheng, Xin Ning, Jun Zhou, and Lin Gu. Talkinggaussian: Structure-persistent 3d talking head synthesis via gaussian splatting. In European Conference on Computer Vision, pp. 127–145. Springer, 2024.

Jiahe Li, Jiawei Zhang, Xiao Bai, Jin Zheng, Jun Zhou, and Lin Gu. Instag: Learning personalized 3d talking head from few-second video. In Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 10690–10700, 2025a.

Tianqi Li, Ruobing Zheng, Minghui Yang, Jingdong Chen, and Ming Yang. Ditto: Motion-space diffusion for controllable realtime talking head synthesis. In Proceedings of the 33rd ACM International Conference on Multimedia, pp. 9704–9713, 2025b.

Chumeng Liang and Xiaoyu Wu. Mist: Towards improved adversarial examples for diffusion models. arXiv preprint arXiv:2305.12683, 2023.

Chumeng Liang, Xiaoyu Wu, Yang Hua, Jiaru Zhang, Yiming Xue, Tao Song, Zhengui Xue, Ruhui Ma, and Haibing Guan. Adversarial example does good: Preventing painting imitation from diffusion models via adversarial examples. arXiv preprint arXiv:2302.04578, 2023.

Pranay Manocha, Zeyu Jin, Richard Zhang, and Adam Finkelstein. Cdpam: Contrastive learning for perceptual audio similarity. In ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 196–200. IEEE, 2021.

Gal Namdar, Galia Avidan, and Tzvi Ganel. Effects of configural processing on the perceptual spatial resolution for face features. Cortex, 72:115–123, 2015.

Ziqiao Peng, Wentao Hu, Yue Shi, Xiangyu Zhu, Xiaomei Zhang, Hao Zhao, Jun He, Hongyan Liu, and Zhaoxin Fan. Synctalk: The devil is in the synchronization for talking head synthesis. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 666–676, 2024.

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.

Yao Qin, Nicholas Carlini, Garrison Cottrell, Ian Goodfellow, and Colin Raffel. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. In International conference on machine learning, pp. 5231–5240. PMLR, 2019.

Zuomin Qu, Zuping Xi, Wei Lu, Xiangyang Luo, Qian Wang, and Bin Li. Df-rap: A robust adversarial perturbation for defending against deepfakes in real-world social network scenarios. IEEE Transactions on Information Forensics and Security, 19:3943–3957, 2024.

Hadi Salman, Alaa Khaddaj, Guillaume Leclerc, Andrew Ilyas, and Aleksander Madry. Raising the cost of malicious ai-powered image editing. In International Conference on Machine Learning, pp. 29894–29918. PMLR, 2023.

Lea Schönherr, Katharina Kohls, Steffen Zeiler, Thorsten Holz, and Dorothea Kolossa. Adversarial attacks against automatic speech recognition systems via psychoacoustic hiding. arXiv preprint arXiv:1808.05665, 2018.

Rui-qing Sun, Xingshan Yao, Tian Lan, Hui-Yang Zhao, Jia-Ling Shi, Chen-Hao Cui, Zhijing Wu, Chen Yang, and Xian-Ling Mao. Efficient and robust video defense framework against 3d-field personalized talking face. arXiv preprint arXiv:2512.21019, 2025.

Ruiqing Sun, Xingshan Yao, Zhijing Wu, Tian Lan, Chenhao Cui, Huiyang Zhao, Jialing Shi, Chen Yang, and Xianling Mao. Do protective perturbations really protect portrait privacy under real-world image transformations? arXiv preprint arXiv:2604.23688, 2026.

James W Tanaka and Joseph A Sengco. Features and their configuration in face recognition. Memory & cognition, 25(5):583–592, 1997.

Jiaxiang Tang, Kaisiyuan Wang, Hang Zhou, Xiaokang Chen, Dongliang He, Tianshu Hu, Jingtuo Liu, Ziwei Liu, Gang Zeng, and Jingdong Wang. Real-time neural radiance talking portrait synthesis via audio-spatial decomposition. International Journal of Computer Vision, 133(9): 6362–6373, 2025.

Hanhui Wang, Yihua Zhang, Ruizheng Bai, Yue Zhao, Sijia Liu, and Zhengzhong Tu. Edit away and my face will not stay: Personal biometric defense against malicious generative editing. In Proceedings of the Computer Vision and Pattern Recognition Conference, pp. 23806–23816, 2025.

Mingwang Xu, Hui Li, Qingkun Su, Hanlin Shang, Liwei Zhang, Ce Liu, Jingdong Wang, Yao Yao, and Siyu Zhu. Hallo: Hierarchical audio-driven visual synthesis for portrait image animation. arXiv preprint arXiv:2406.08801, 2024.

Haotian Xue, Chumeng Liang, Xiaoyu Wu, and Yongxin Chen. Toward effective protection against diffusion-based mimicry through score distillation. In The Twelfth International Conference on Learning Representations, 2023.

Zhenhui Ye, Ziyue Jiang, Yi Ren, Jinglin Liu, Jinzheng He, and Zhou Zhao. Geneface: Generalized and high-fidelity audio-driven 3d talking face synthesis. arXiv preprint arXiv:2301.13430, 2023.

Zhenhui Ye, Tianyun Zhong, Yi Ren, Jiaqi Yang, Weichuang Li, Jiawei Huang, Ziyue Jiang, Jinzheng He, Rongjie Huang, Jinglin Liu, et al. Real3d-portrait: One-shot realistic 3d talking portrait synthesis. arXiv preprint arXiv:2401.08503, 2024.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In Advances in Neural Information Processing Systems, volume 33, 2020.

Yue Zhang, LIU Minhao, Zhaokang Chen, Bin Wu, Chao Zhan, Yingjie He, JUNXIN HUANG, Wenjiang Zhou, et al. Musetalk: Real-time high quality lip synchronization with latent space inpainting. 2024.

Zhimeng Zhang, Lincheng Li, Yu Ding, and Changjie Fan. Flow-guided one-shot talking face generation with a high-resolution audio-visual dataset. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 3661–3670, 2021.

Eberhard Zwicker and Ernst Terhardt. Analytical expressions for critical-band rate and critical bandwidth as a function of frequency. The Journal of the Acoustical Society of America, 68(5): 1523–1525, 1980.

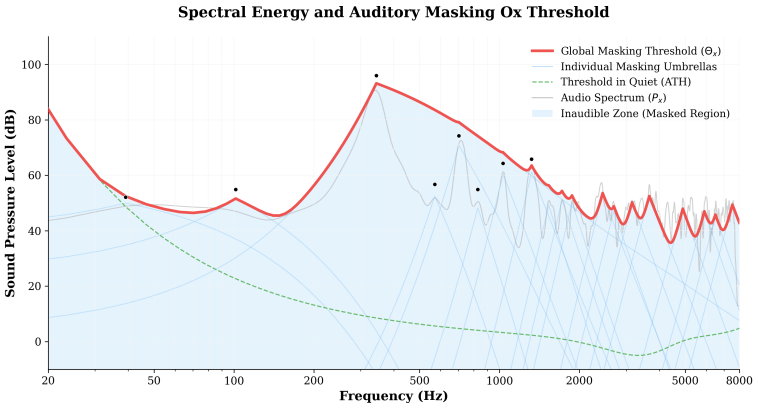


Figure 4: Illustration of psychoacoustic masking used in our defense. Strong spectral components raise local masking thresholds, allowing defensive perturbations below the threshold to remain less perceptible.

A Psychoacoustic Masking Threshold Computation

This section details the computation of the global masking threshold $\Theta_{\mathbf{x}}(t, f)$ introduced in Sec. 3.2. The formulation follows established psychoacoustic masking models (Qin et al., 2019; Schönherr et al., 2018), where strong spectral components act as maskers and increase the perceptual threshold of nearby frequency components. As illustrated in Fig. 4, this effect creates masked regions in which weaker perturbations are less perceptible. The computation consists of spectral normalization, perceptual frequency mapping, masker detection, individual masking estimation, and global threshold aggregation.

Time Frequency Representation and PSD Normalization. Given a source audio signal $\mathbf{x}$, a short time Fourier transform (STFT) is first applied to obtain its complex spectrum $X(t, k)$ at time frame t and frequency bin k. The power spectral density (PSD) in dB is computed as

$$P_{\mathbf{x}}(t, k) = 10 \log_{10} \left| \frac{1}{N} X(t, k) \right|^2, \tag{12}$$

where N denotes the STFT window length. Following prior psychoacoustic audio optimization (Qin et al., 2019), the PSD of each frame is normalized to a common reference level as

$$\bar{P}_{\mathbf{x}}(t, k) = P_{\text{top}} - \max_{k'} P_{\mathbf{x}}(t, k') + P_{\mathbf{x}}(t, k), \tag{13}$$

where P_{top} denotes the reference sound pressure level. For frequency bin k, its corresponding physical frequency is

$$f_k = \frac{k}{N} f_s, \qquad 0 \le f_k \le \frac{f_s}{2}, \tag{14}$$

where f_s denotes the audio sampling rate.

The perturbation PSD is normalized using the same clean audio reference,

$$\bar{P}_{\delta}(t, k) = P_{\text{top}} - \max_{k'} P_{\mathbf{x}}(t, k') + P_{\delta}(t, k), \tag{15}$$

which places $\bar{P}_{\delta}(t, k)$ and the masking threshold derived below on the same normalized dB scale.

Perceptual Frequency Mapping and Hearing Threshold. Human auditory frequency resolution is nonuniform and can be characterized by critical bands on the Bark scale (Zwicker & Terhardt, 1980; Fastl & Zwicker, 2006). A frequency f in Hz is mapped to its Bark position as

$$z(f) = 13 \arctan(0.00076 f) + 3.5 \arctan\left[\left(\frac{f}{7500} \right)^2 \right]. \tag{16}$$

The absolute threshold of hearing provides the minimum audible level in a quiet environment and is approximated as

$$T_q(f) = 3.64\left(\frac{f}{1000}\right)^{-0.8} - 6.5\exp\left[-0.6\left(\frac{f}{1000} - 3.3\right)^2\right] + 10^{-3}\left(\frac{f}{1000}\right)^4. \tag{17}$$

Only frequency components within the effective human hearing range and above $T_q(f)$ are considered as candidate maskers (Qin et al., 2019).

Masker Detection and Individual Masking Thresholds. For each time frame t, candidate maskers are selected from $\bar{P}_{\mathbf{x}}(t,k)$. A spectral component is retained as a masker if it is a local spectral maximum, exceeds the absolute threshold of hearing, and has the highest PSD within a 0.5 Bark neighborhood (Qin et al., 2019). The resulting set of maskers at frame t is denoted by $\mathcal{M}_t$.

For a masker centered at frequency bin k_i, its spectral level is smoothed with its two neighboring bins in the intensity domain as

$$\bar{P}^m_{\mathbf{x}}(t,k_i) = 10\log_{10}\left(10^{\bar{P}_{\mathbf{x}}(t,k_i-1)/10} + 10^{\bar{P}_{\mathbf{x}}(t,k_i)/10} + 10^{\bar{P}_{\mathbf{x}}(t,k_i+1)/10}\right). \tag{18}$$

Let $z_i = z(f_i)$ denote the Bark position of masker i, and let

$$\Delta z = z(f) - z_i. \tag{19}$$

The masking influence is modeled using the two slope spreading function (Qin et al., 2019)

$$SF_i(\Delta z) = \begin{cases} 27\Delta z, & \Delta z \le 0, \\ \left[-27 + 0.37\max\left(\bar{P}^m_{\mathbf{x}}(t,k_i) - 40, 0\right)\right]\Delta z, & \Delta z > 0. \end{cases} \tag{20}$$

The individual masking threshold induced by masker i at frequency f is then computed as

$$T_i(t,f) = \bar{P}^m_{\mathbf{x}}(t,k_i) - 6.025 - 0.275 z_i + SF_i(\Delta z). \tag{21}$$

Global Masking Threshold. The absolute threshold of hearing and all individual masking thresholds are combined in the intensity domain to obtain the global masking threshold

$$\Theta_{\mathbf{x}}(t,f) = 10\log_{10}\left(10^{T_q(f)/10} + \sum_{i\in\mathcal{M}_t} 10^{T_i(t,f)/10}\right). \tag{22}$$

The resulting $\Theta_{\mathbf{x}}(t,f)$ characterizes the estimated perceptual boundary of the source audio at each time frequency location. As described in Sec. 4.1, perturbation components exceeding the margin adjusted threshold $\Theta_{\mathbf{x}}(t,f) - m$ are penalized by $\mathcal{L}_{\text{mask}}$.

B Details of Continual Optimization

This section provides implementation details of the continual multi encoder optimization and the second stage perceptual refinement used in VoxGuard.

Adaptive Objective Balancing. As described in Sec. 4.2, each encoder specific redirection loss is first normalized by its clean source to target distance,

$$L_i^{\text{rel}} = \frac{L_i(\tilde{x}, x^{\text{tar}})}{L_i(x, x^{\text{tar}})}. \tag{23}$$

For the active encoder set $\mathcal{A}_k$, the weight $w_{i,k}$ in Eq. (9) is recomputed at every optimization step according to

$$w_{i,k} = \frac{1-\rho}{|\mathcal{A}_k|} + \rho\frac{\exp\left(\tau L_i^{\text{rel}}\right)}{\sum_{j\in\mathcal{A}_k}\exp\left(\tau L_j^{\text{rel}}\right)}, \qquad i \in \mathcal{A}_k, \tag{24}$$

where $\rho = 0.5$ and $\tau = 10$ in all experiments. The first term provides a uniform contribution from all active encoders, while the second assigns greater weight to encoders with larger remaining

relative loss, preventing poorly optimized representations from being dominated by objectives that have already undergone stronger redirection.

Objective weighting and gradient scale balancing are performed separately. The perturbation is parameterized as $\delta = \epsilon \tanh(\theta)$, and θ is directly optimized. For encoder i, we compute

$$g_i = \nabla_\theta L_i^{\mathrm{rel}}. \tag{25}$$

Because heterogeneous encoders can produce substantially different gradient magnitudes, their gradients are matched to a common RMS scale before aggregation. Up to a shared positive scale, the encoder update direction is

$$g_k^{\mathrm{enc}} = \sum_{i \in \mathcal{A}_k} w_{i,k} \frac{g_i}{\mathrm{RMS}(g_i)}. \tag{26}$$

Thus, $w_{i,k}$ balances optimization progress across encoders, whereas RMS matching prevents differences in gradient magnitude from determining their contribution.

Continual Encoder Schedule. We use three surrogate audio representations in the fixed order AVE, Esperanto, and HuBERT. The first optimization stage contains three consecutive blocks with equal budgets of 80 steps per block. The first block optimizes AVE only. The second introduces Esperanto while retaining the established AVE redirection. The third introduces HuBERT while retaining the redirection established for AVE and Esperanto. Accordingly, the active encoder sets are

$$\{\mathrm{AVE}\} \to \{\mathrm{AVE}, \mathrm{Esperanto}\} \to \{\mathrm{AVE}, \mathrm{Esperanto}, \mathrm{HuBERT}\}. \tag{27}$$

Stage 1 therefore contains 240 optimization steps in total.

For a previously optimized encoder j, we define its retention ceiling at the beginning of block k as

$$c_{j,k} = \min\left\{(1+\gamma) b_{j,k}, 1-\eta\right\}, \tag{28}$$

where $b_{j,k}$ is its relative loss at block entry, $\gamma = 0.05$ is the allowed forgetting ratio, and $\eta = 10^{-6}$ prevents a retained encoder from regressing beyond its clean baseline. A candidate update is feasible only if

$$L_j^{\mathrm{rel}}(\tilde{x}) \le c_{j,k}, \qquad \forall j \in \mathcal{P}_k. \tag{29}$$

Conflict Projection. Let g_k denote the aggregated update direction and g_j the RMS matched gradient of a protected encoder $j \in \mathcal{P}_k$. When their inner product is negative, updating along $-g_k$ would increase the protected objective to first order. We therefore remove the conflicting component as

$$g_k \leftarrow g_k - \frac{g_k^\top g_j}{\|g_j\|_2^2} g_j, \qquad g_k^\top g_j < 0. \tag{30}$$

Projection is successively applied to the protected objectives for at most three passes. The resulting direction is then rescaled to preserve the RMS magnitude before projection, suppressing conflicting components without unnecessarily reducing the overall update strength.

Retention Feasible Backtracking. The perturbation budget is enforced through the bounded parameterization

$$\delta = \epsilon \tanh(\theta), \tag{31}$$

with $\epsilon = 0.02$. For the resulting update direction g_k with respect to θ, a candidate at backtracking iteration m is

$$\theta^{(m)} = \theta - \alpha 2^{-m} \frac{g_k}{\mathrm{RMS}(g_k)}, \tag{32}$$

where $\alpha = 0.01$ and $m = 0, \ldots, 12$. During the first optimization stage, a candidate is accepted only when the relative loss of the newly introduced encoder decreases and all protected encoder losses satisfy Eq. equation 28. Otherwise, the step size is successively halved until a feasible update is obtained. If no feasible candidate is found, the current iterate is retained.

Perceptual Refinement. The second stage starts from the perturbation obtained after Stage 1 and repeats the same AVE, Esperanto, and HuBERT schedule with 80 steps per block, giving another 240 optimization steps. Psychoacoustic and temporal constraints are enabled throughout this stage while the encoder redirection update remains active.

Let

$$L_{\text{reg}} = \lambda_{\text{mask}} L_{\text{mask}} + R(\delta), \tag{33}$$

where

$$R(\delta) = \lambda_0 \frac{1}{N} \|\delta\|_2^2 + \lambda_1 \frac{1}{N-1} \|\nabla_t \delta\|_2^2 + \lambda_2 \frac{1}{N-2} \|\nabla_t^2 \delta\|_2^2. \tag{34}$$

The regularization gradient is computed directly with respect to the optimization variable θ,

$$g_k^{\text{reg}} = \nabla_\theta L_{\text{reg}}, \tag{35}$$

and combined with the RMS balanced encoder direction as

$$g_k^{(2)} = g_k^{\text{enc}} + g_k^{\text{reg}}. \tag{36}$$

We use

$$\lambda_{\text{mask}} = 5, \qquad \lambda_0 = 5 \times 10^4, \qquad \lambda_1 = \lambda_2 = 100, \tag{37}$$

with a psychoacoustic safety margin of 3 dB. The three terms in $R(\delta)$ respectively constrain perturbation energy, first order temporal variation, and second order temporal variation.

The combined direction $g_k^{(2)}$ is subsequently processed by the same conflict projection used in Stage 1. Retention feasible backtracking is then applied, with the encoder weighting coefficients and RMS matching scales fixed within each line search. A candidate is accepted only when the resulting refinement criterion decreases while all protected encoder objectives remain within their retention ceilings. This preserves the established representation shifts while progressively reducing perceptual distortion. The second stage starts from the perturbation obtained after Stage 1 and repeats the same AVE, Esperanto, and HuBERT schedule with 80 steps per block, giving another 240 optimization steps. Psychoacoustic and temporal constraints are enabled throughout this stage while the encoder redirection objective remains active. The refinement objective is

$$J_k^{(2)} = L_k^{\text{enc}} + \lambda_{\text{mask}} L_{\text{mask}} + R(\delta), \tag{38}$$

where

$$R(\delta) = \lambda_0 \frac{1}{N} \|\delta\|_2^2 + \lambda_1 \frac{1}{N-1} \|\nabla_t \delta\|_2^2 + \lambda_2 \frac{1}{N-2} \|\nabla_t^2 \delta\|_2^2. \tag{39}$$

We use

$$\lambda_{\text{mask}} = 5, \qquad \lambda_0 = 5 \times 10^4, \qquad \lambda_1 = \lambda_2 = 100, \tag{40}$$

with a psychoacoustic safety margin of 3 dB. The three regularization terms respectively constrain perturbation energy, first order temporal variation, and second order temporal variation.

During refinement, gradients of the complete objective are computed directly with respect to θ. The same conflict projection and retention feasible backtracking are then applied. A candidate update is accepted only when the regularized refinement objective decreases while all retained encoder objectives remain within their corresponding retention ceilings. This prevents perceptual refinement from erasing the representation shifts established in the first stage.

Overall, VoxGuard uses 240 redirection steps in Stage 1 and 240 refinement steps in Stage 2, corresponding to 80 optimization steps for each encoder block in each stage.

C Additional Experimental Details

C.1 Generation Models and Audio Encoders

We evaluate our defense on two representative personalized audio-driven 3D talking face generation models with different underlying 3D representations. SyncTalk (Peng et al., 2024) is a NeRF-based method, whereas InsTaG (Li et al., 2025a) is based on 3D Gaussian Splatting. Evaluating both models allows us to examine whether the proposed audio-side defense remains effective across different personalized 3D rendering paradigms.

For the audio front end, we consider four encoders supported by modern talking-face generation pipelines: AVE (Peng et al., 2024), Esperanto based on the wav2vec 2.0 family (Baevski et al., 2020; Conneau et al., 2020), HuBERT (Hsu et al., 2021), and DeepSpeech (Amodei et al., 2016). These encoders span heterogeneous speech representations and include both convolutional and Transformer-based architectures.

During perturbation optimization, AVE, Esperanto, and HuBERT are used as surrogate encoders. DeepSpeech is completely excluded from optimization and is only introduced at evaluation time. It therefore provides an unseen audio representation for assessing black-box transferability.

C.2 Baselines and Datasets

We adopt Talk9 (Li et al., 2025b) as the primary evaluation benchmark. To further examine generalization across identities, 40 identities are randomly sampled from the HDTF dataset (Zhang et al., 2021). Because personalized 3D TFG requires an independent personalization process for each identity, this subset provides substantially broader identity coverage while keeping the computational cost tractable.

We compare our method with representative visual-domain proactive defenses, including AdvDM (Liang et al., 2023), PhotoGuard (Salman et al., 2023), the SDS family (Xue et al., 2023), and Silencer (Gan et al., 2025).We do not include the recent defense for personalized 3D-field TFG (**?**) because no official implementation was publicly available at the time of submission, and a nonofficial reimplementation could introduce implementation-dependent bias into the comparison. These methods protect portrait media by perturbing visual content, whereas our method modifies only the audio track. This comparison is therefore intended to examine whether effective protection can be achieved without directly modifying the facial appearance of the released reference video.

C.3 Evaluation Metrics

We evaluate VoxGuard from three complementary perspectives: audio-visual synchronization, generated video quality, and protected audio quality.

Audio-Visual Synchronization. Sync-C and Sync-D computed by SyncNet (Chung & Zisserman, 2016) are used to measure the correspondence between the generated facial motion and the driving audio. Sync-D measures the distance between audio and visual embeddings, whereas Sync-C reflects synchronization confidence. Accordingly, lower Sync-C and higher Sync-D indicate weaker audio-visual synchronization and therefore stronger defensive disruption. Undefended results are reported as a direct reference.

Generated Video Quality. We further report PSNR, SSIM, and FID (Heusel et al., 2017) to quantify the degradation of the generated visual content. Lower PSNR and SSIM indicate reduced reconstruction similarity, while higher FID indicates a larger distributional discrepancy from the reference content. Thus, lower PSNR and SSIM together with higher FID correspond to stronger degradation of the generated video.

Protected Audio Quality. We use CDPAM (**?**) and signal-to-noise ratio (SNR) to quantify the distortion introduced into the protected audio. CDPAM measures perceptual audio distance using learned representations, with lower values indicating greater perceptual similarity to the clean audio. SNR measures the relative energy of the clean signal and the introduced perturbation,

$$\mathrm{SNR} = 10 \log_{10} \frac{\|x\|_2^2}{\|\tilde{x} - x\|_2^2}. \tag{41}$$

Higher SNR therefore indicates lower signal-level distortion. Together, lower CDPAM and higher SNR indicate better preservation of the original audio quality.

Protected Visual Quality. For visual-domain defenses, we further measure the distortion introduced into the released reference frames using PSNR, SSIM, and FID. Higher PSNR and SSIM and lower FID indicate better preservation of the original visual content. Since VoxGuard does not modify visual frames, its protected frames are pixel-wise identical to the clean reference frames.

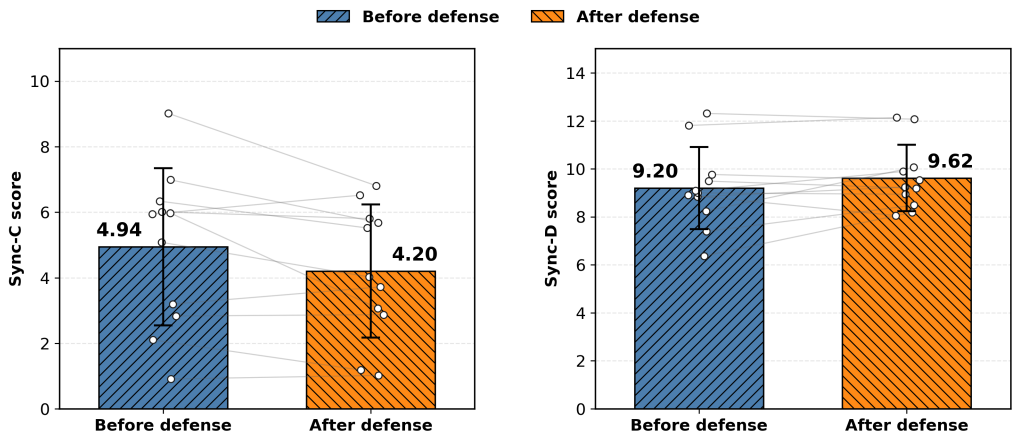


Figure 5: Cross-pipeline evaluation with identity-specific MuseTalk 1.5 adaptation on Talk9. Connected circles show paired results for the same identity trained with clean and protected reference videos. Bars and error bars denote the mean and standard deviation across identities.

D Cross-Pipeline Evaluation with MuseTalk 1.5

Experimental setup. We further examine whether our defense interferes with identity-specific facial motion acquisition beyond the personalized 3D TFG pipelines used in our primary evaluation. We adopt the pretrained MuseTalk 1.5 model (Zhang et al., 2024), whose generation architecture differs substantially from SyncTalk and InsTaG and whose Whisper-based audio representation is not used during perturbation optimization. For each Talk9 identity, we initialize from the same pretrained checkpoint and fully fine-tune the model using either the clean or protected reference video. All other training settings and the evaluation audio are kept identical between the paired runs. We measure audio-visual synchronization using Sync-C and Sync-D (Chung & Zisserman, 2016), where lower Sync-C and higher Sync-D indicate weaker synchronization.

Results. As shown in Fig. 5, adaptation with protected reference videos consistently shifts synchronization in the defensive direction, reducing the average Sync-C from 4.945 to 4.202 while increasing Sync-D from 9.202 to 9.619. The paired identity-level results further show that this degradation is not driven by a small subset of identities. These results provide complementary evidence that our defense disrupts the acquisition of identity-specific audio-to-facial-motion correspondence even when both the generation pipeline and audio representation differ from those considered during perturbation optimization. Since MuseTalk lies outside the personalized 3D-field setting targeted by this work, we use this experiment as evidence of cross-pipeline transfer rather than as part of the primary benchmark.

E Human Perceptual Evaluation Details

Participants and evaluation protocol. Fourteen student volunteers participated anonymously in the perceptual study. Participation was voluntary and uncompensated, and all participants completed the full evaluation with no responses excluded. No demographic information was collected. We use all nine identities from Talk9 and select one 10 s reference segment for each identity. The same source segment is used across all evaluated conditions, including Clean, Ours, Silencer, AdvDM($-$), SDS($-$), SDST($-$), PhotoGuard, Mist, SDS($+$), and AdvDM($+$). This results in 90 evaluation samples in total.

Each participant evaluated all 90 samples. The samples were presented individually in a randomized order, and the corresponding defense method was hidden from the participants. Each sample could be played once and replay was disabled. Participants rated the same sample along three dimensions in the fixed order of Overall, Visual, and Audio, and submitted the three scores together. Over-

Table 4: Unified five point scale used in the human perceptual evaluation.

Score	Rating criterion
1	**Unacceptable.** Clear visual or auditory degradation is present and the sample is unacceptable.
2	**Poor.** Clear perceptual degradation is present and the overall experience is poor.
3	**Marginally acceptable.** The sample remains viewable, but perceptual degradation is evident.
4	**Acceptable.** The overall experience is acceptable with only minor perceptual degradation.
5	**Fully acceptable.** Almost no perceptible quality degradation is observed.

all measures the perceptual quality of the audiovisual sample as a whole, while Visual and Audio separately assess the perceived quality of the visual and audio components.

Participants completed the study on a computer using headphones in a quiet environment. Before the formal evaluation, several calibration clips were provided to adjust the playback volume to a comfortable level. The volume was then kept fixed throughout the evaluation. The interface required all three ratings before submission and prevented repeated participation.

Rating scale. All three dimensions use the same five point perceptual quality scale. The complete rating criteria are reported in Table 4.

Score aggregation. Each method receives nine identity level evaluations from each of the 14 participants, resulting in 126 ratings per method for each evaluation dimension. The reported Overall, Visual, and Audio scores are obtained by directly averaging all participant and identity ratings for the corresponding method. Across the complete study, this yields 1,260 sample evaluations and 3,780 individual quality ratings.

Interpretation of modality specific ratings. We observe small cases where an unchanged modality receives a slightly higher subjective score than Clean. For example, our method leaves all visual frames unchanged but obtains a Visual score of 4.74 compared with 4.65 for Clean, while several visual defenses similarly receive slightly higher Audio scores despite leaving the audio unchanged. Prior studies have shown that subjective audio and visual quality judgments are not perceptually independent and can be affected by cross modal interaction, attention, and evaluation context (Belmudez & Möller, 2013). We therefore regard these small differences as subjective perceptual variation rather than evidence that the unchanged modality is enhanced.

F Hyperparameter Analysis

We examine the sensitivity of our defense to two primary optimization parameters, the perturbation budget ϵ and the number of optimization steps. We vary one parameter at a time while keeping the other at its default value. The default configuration uses $\epsilon = 0.02$ and 80 optimization steps per encoder block. To make metrics with different directions directly comparable, Fig. 6 reports relative improvement over the default configuration, where positive values consistently indicate better performance. Lower Sync-C and CDPAM and higher Sync-D and SNR are therefore treated as improvements. The curves report the mean across Talk9 identities, while the shaded regions denote the corresponding standard deviation.

Perturbation Budget. As shown in Fig. 6(a), reducing ϵ improves perceptual quality but substantially weakens defense effectiveness. At $\epsilon = 0.005$ and 0.01, both SNR and CDPAM improve relative to the default setting, while Sync-C and Sync-D indicate clearly weaker synchronization disruption. Increasing ϵ beyond 0.02 produces the opposite behavior. Synchronization disruption becomes stronger, particularly in terms of Sync-C, while SNR decreases noticeably. In contrast, CDPAM changes only mildly at larger perturbation budgets, suggesting that psychoacoustic masking continues to limit perceptually salient distortion even as the perturbation magnitude increases. These results reveal a clear balance between representation disruption and audio quality.

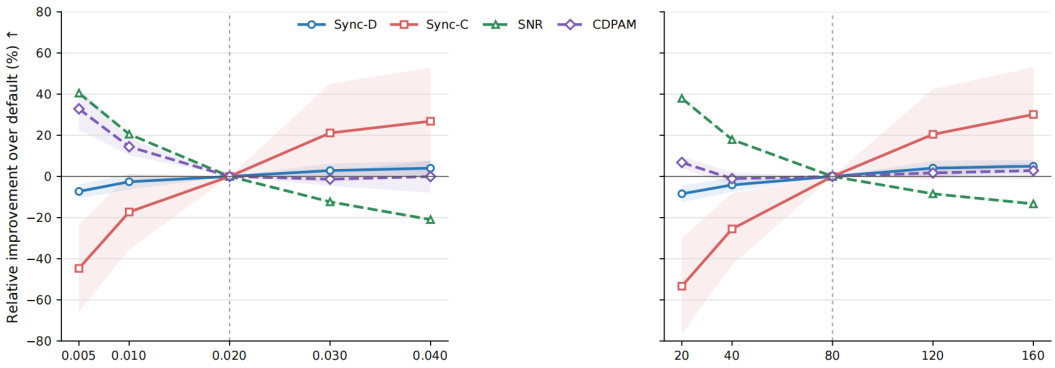


Figure 6: Sensitivity to the perturbation budget ϵ and the number of optimization steps on Talk9. Each parameter is varied independently while the other is fixed at its default value of $\epsilon = 0.02$ and 80 steps per encoder block in each stage. Results are reported as relative improvement over the default configuration. Positive values indicate better performance, where lower Sync-C and CDPAM and higher Sync-D and SNR are preferred. Curves show the mean across identities and shaded regions show the standard deviation.

Optimization Steps. A similar pattern is observed in Fig. 6(b). Reducing the number of optimization steps to 20 or 40 preserves higher SNR but does not establish sufficiently strong synchronization disruption. Increasing the number of steps to 120 or 160 further improves Sync-C and produces moderate gains in Sync-D, while SNR decreases. CDPAM remains close to the default setting, indicating that the additional perturbation energy does not translate proportionally into perceptually salient degradation. The larger variation in Sync-C also suggests that sensitivity to aggressive optimization differs across identities.

Overall, $\epsilon = 0.02$ and 80 optimization steps per encoder block in each stage provide a balanced configuration that achieves effective synchronization disruption without requiring a substantially larger perturbation budget or longer optimization. We therefore adopt this configuration as the default setting throughout the experiments.

G Additional Visual Comparison

To further illustrate the visual impact of different proactive defense methods, we provide additional qualitative comparisons of the protected reference videos. As shown in Fig. 7, visual-domain defenses introduce varying degrees of visible perturbations and local artifacts on facial regions. In contrast, our audio-side defense does not modify the visual frames, and therefore preserves the original facial appearance.

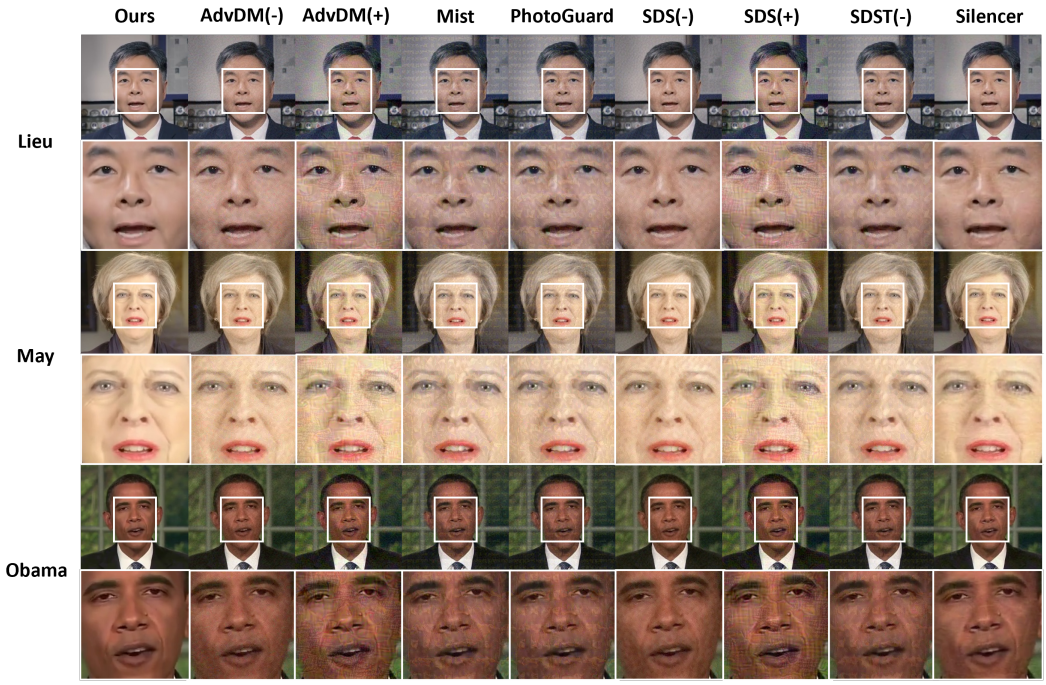


Figure 7: Additional visual comparison of protected reference videos across different identities and defense methods. For each identity, the upper row shows the protected reference frames, while the lower row presents enlarged views of the marked facial regions to highlight local visual distortions.