SYNCEDIT: SEMANTICALLY SELECTIVE LIP-SYNC DISRUPTION FOR TALKING-HEAD GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio-driven talking-head systems can be misused to fabricate convincing videos in which a target identity appears to deliver harmful speech. Existing defenses perturb the supplied portrait or audio, requiring per-input optimization without considering speech semantics. We propose SYNCEDIT, a model-level defense that retrains only the audio encoder while keeping downstream modules unchanged. SYNCEDIT disrupts audio–lip synchronization for harmful speech while preserving generation for semantically benign speech both with and without harmful keywords, without requiring speech-category labels or external classifiers at inference time. Direct joint training of these objectives stalls because gradients for harmful-speech disruption interfere with those for benign-speech preservation. We address this conflict through two-stage optimization and coupling-based layer selection. Experiments on four heterogeneous talking-head systems show that SYNCEDIT induces substantially larger representation and synchronization changes for harmful speech than for either benign category, while preserving identity and visual quality. A human study further confirms that SYNCEDIT reduces perceived mouth match and credibility for harmful videos while leaving both benign categories nearly unchanged. Ablations validate the proposed training design, and further experiments demonstrate generalization to unseen speech distributions and partial resistance to a white-box adaptive attack.

1 INTRODUCTION

Audio-driven talking-head generation synthesizes facial motion or portrait videos from driving speech and an identity reference. Recent systems produce realistic, lip-synchronized outputs for applications such as digital avatars, virtual presenters, and video dubbing Xu et al. (2024); Cheng et al. (2025); Tan et al. (2025); Fan et al. (2024). However, they can also be misused to make a target identity appear to deliver harmful speech, thereby inciting hatred, discrediting individuals or groups, manipulating public opinion, and causing reputational harm. Because the credibility of such videos depends heavily on audio–lip synchronization, weakening this coupling provides a direct means of defense.

Existing defenses specifically designed for talking-head misuse disrupt facial motion and lip synchronization through input perturbations: Silencer modifies the reference portrait Gan et al. (2025), Sun et al. perturb the driving audio Sun et al. (2026), and SyncBreaker perturbs both inputs Zhang et al. (2026). They leave the model unchanged, require per-input optimization, and depend on the protected inputs being used. They also suppress speech-driven animation without considering speech semantics and therefore cannot selectively disrupt harmful speech while preserving legitimate use.

To address this gap, we propose SYNCEDIT (Figure 1), a model-level defense that explicitly distinguishes three speech categories: harmful speech ($\mathcal{H}$), keyword-benign speech that contains harmful keywords but remains semantically benign ($\mathcal{KB}$), and ordinary benign speech without such keywords ($\mathcal{B}$). To our knowledge, SYNCEDIT is the first talking-head defense to consider the $\mathcal{KB}$ category, which is essential for preventing the defense from reacting indiscriminately to harmful keywords. Its objective is to disrupt audio–lip synchronization for $\mathcal{H}$ while preserving the original generation behavior for $\mathcal{KB}$ and $\mathcal{B}$. SYNCEDIT realizes this selective behavior by retraining only the audio encoder while keeping all other modules unchanged. During offline training, it compares the generator-facing audio conditions produced by the original and trainable encoders. It encourages these conditions to deviate for $\mathcal{H}$ while preserving them for $\mathcal{KB}$ and $\mathcal{B}$.

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.

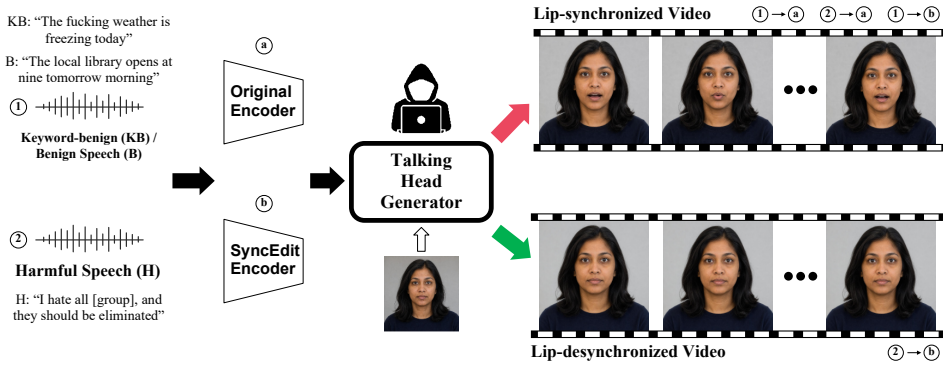


Figure 1: Illustration of SYNCEDIT. SYNCEDIT edits the audio encoder while keeping other modules fixed. For harmful speech, the defended system produces a visible harmful audio–lip mismatch. For benign hard negatives and ordinary benign speech, output changes remain much smaller.

A key technical challenge is that directly optimizing the three category-specific objectives from the original encoder fails to produce the intended behavior. Our analysis reveals a structured gradient conflict: updates that increase the representation shift for $\mathcal{H}$ interfere with those that preserve $\mathcal{KB}$ and $\mathcal{B}$, causing optimization to stall. SYNCEDIT addresses this conflict through two-stage training. Stage 1 jointly optimizes $\mathcal{H}$ and $\mathcal{B}$, whose conflict is comparatively weaker, to establish a controlled harmful-speech shift. Stage 2 starts from the Stage 1 checkpoint and introduces $\mathcal{KB}$, after the harmful–benign separation has already been established and the gradient conflict has substantially weakened. Stage 2 further updates only encoder layers in which $\mathcal{H}$ and $\mathcal{KB}$ representations exhibit sufficiently low coupling, thereby preserving $\mathcal{KB}$ without erasing the harmful-speech shift learned in Stage 1.

We evaluate SYNCEDIT on four heterogeneous talking-head systems, covering four audio encoders and both RGB-video and geometry-based generation pipelines. Across all systems, SYNCEDIT induces substantially larger representation and synchronization changes for $\mathcal{H}$ than for $\mathcal{KB}$ and $\mathcal{B}$, while preserving identity and visual quality. A human study further shows that it reduces the mouth-matching and video credibility rating of harmful videos by 1.20 and 1.06 points on a five-point scale, whereas the changes for $\mathcal{KB}$ and $\mathcal{B}$ are at most 0.07 points. Ablation experiments confirm the importance and contribution of two-stage training and selected-layer optimization. The learned behavior also generalizes to an unseen speech synthesizer and genuine human speech. Under a white-box adaptive attack, audio–lip synchronization is partially restored but remains worse than that of the undefended system, showing that the attack weakens but does not fully bypass the defense.

Our main contributions are as follows:

- We formulate category-selective defense for talking-head generation, including the previously overlooked $\mathcal{KB}$ hard category, and propose SYNCEDIT, which internalizes this behavior by retraining only the audio encoder.
- We identify gradient conflict as a key obstacle to selective encoder retraining and address it through two-stage optimization and coupling-based layer selection.
- We demonstrate the effectiveness, utility, and robustness of SYNCEDIT across four talking-head systems through automated metrics, a human study, ablation experiments, distribution-shift evaluation, and a white-box adaptive attack.

2 RELATED WORKS

Audio-driven talking-head generation. Audio-driven talking-head systems generate facial motion or portrait videos synchronized with driving speech while preserving an identity specified by a reference input. Video-based systems, including Hallo Xu et al. (2024), DAWN Cheng et al. (2025), and DICE-Talk Tan et al. (2025), synthesize portrait videos directly, whereas geometry-based systems such as UniTalker Fan et al. (2024) predict 3D facial motion or animation.

Given driving speech x and an identity reference r, the generation process can be abstracted as

$$y = G\left(r, A(E(x))\right), \tag{1}$$

where E is an audio encoder that extracts a sequence of speech representations; the adaptor A maps the representations to the conditioning signals expected by the downstream generator; and G denotes the remaining generation pipeline, including video synthesis, motion prediction, and rendering, depending on the specific system.

Input-side safeguards for talking-head misuse. Input-side safeguards include *input detection* and *input perturbation*. Although not designed specifically for talking-head systems, audio hate-speech detectors Gupta et al. (2022); Ghosh et al. (2022); An et al. (2024); Costa-jussà et al. (2024) could be used to screen driving speech before generation. Perturbation methods modify supplied media to disrupt generation by suppressing audio-driven motion, degrading 3D facial animation, or impairing lip synchronization and facial dynamics: Silencer Gan et al. (2025), Sun et al. Sun et al. (2026), and SyncBreaker Zhang et al. (2026) perturb the portrait, driving audio, and both modalities, respectively.

These defenses remain external to the generation model. Detectors introduce inference overhead, may misclassify semantically benign speech containing sensitive words, and can be disabled in self-hosted open-source systems by removing the detector from the inference code. Perturbation methods require input-specific optimization and may be bypassed through purification or input replacement. In contrast, SYNCEDIT embeds selective defense into the audio encoder and requires no external classifier, speech-category label, or per-input perturbation.

Model editing and concept erasure for conditional generation. Post-hoc modification of generative models is commonly studied as *model editing*, with safety-oriented removal often termed *concept erasure* or *concept unlearning*. Existing methods modify either the generation backbone or its conditioning representations. ESD Gandikota et al. (2023) and MACE Lu et al. (2024) edit diffusion-model parameters to erase target concepts, while SafeGen Li et al. (2024) modifies self-attention to suppress sexually explicit visual representations. ReFACT Arad et al. (2024), AdvUnlearn Zhang et al. (2024), and SafeText Hu et al. (2025) edit text encoders to update factual associations or suppress unsafe concepts, whereas SafeRedir Liu et al. (2026) redirects unsafe prompt embeddings at inference time. These methods target text-conditioned image generation. Talking-head defense instead requires controlling temporal audio-to-motion generation while disrupting $\mathcal{H}$ and preserving $\mathcal{KB}$ and $\mathcal{B}$. SYNCEDIT addresses the resulting gradient conflict by retraining only the audio encoder through two-stage training with selected-layer optimization.

3 THREAT MODEL

Attacker's goal. The attacker uses a talking-head generation system to create a credible, lip-synchronized video in which a target identity appears to deliver a specified harmful utterance. The attack succeeds if the generated mouth motion matches the speech closely enough to appear credible. Such videos may incite hatred, discredit individuals or groups, manipulate public opinion, and damage the target's reputation, thereby fostering a toxic and untrustworthy online environment.

Attacker's capability. We consider two attacker settings. A black-box, non-adaptive attacker interacts with the deployed system only through its standard interface, cannot access its internals, and is unaware of SYNCEDIT. A white-box, adaptive attacker has full access to the system and knowledge of SYNCEDIT, and can optimize input-waveform perturbations to evade the defense.

Defender's objective. Input speech belongs to one of three categories. $\mathcal{H}$ denotes semantically harmful utterances containing explicit harmful keywords (e.g., *"I hate all [group], and they should be eliminated"*). $\mathcal{KB}$ denotes utterances containing such keywords that remain semantically benign in context (e.g., *"The fucking weather is freezing today"*). $\mathcal{B}$ denotes ordinary benign utterances without harmful keywords (e.g., *"The local library opens at nine tomorrow morning"*).

The defender designs the selective defense SYNCEDIT to achieve two objectives.

Effectiveness. For $x \in \mathcal{H}$, the generated mouth motion should deviate substantially from the driving speech, producing a clear audio–lip mismatch and reducing video credibility.

Utility. For $x \in \mathcal{KB} \cup \mathcal{B}$, the defended output should remain close to that of the undefended system, preserving generation quality and audio–lip synchronization.

Defender's capability. The defender is the developer or maintainer of the talking-head system and therefore has white-box access to the deployed model and can update its parameters. During offline

training, the defender has access to a labeled speech corpus containing $\mathcal{H}$, $\mathcal{KB}$, and $\mathcal{B}$ utterances. At inference time, however, SYNCEDIT does not require the speech label to select a corresponding processing path. Instead, this selective defensive behavior is internalized in the defended system.

4 SYNCEDIT

4.1 OVERVIEW

As shown in Figure 2, SYNCEDIT is a model-level defense that retrains only the audio encoder while keeping the audio adaptor and downstream generator fixed. This reduces training overhead and avoids modifying the often complex, model-specific generation pipeline. It shifts the generator-facing conditions of $\mathcal{H}$ speech while preserving those of $\mathcal{KB}$ and $\mathcal{B}$ speech.

Because the generator consumes the adaptor output rather than the raw encoder representation, we define the objective at this interface. Applying it directly to the encoder representation yields substantially weaker defense effectiveness, as shown in § 5.4. Let E^0 denote the frozen original encoder and E^θ its trainable copy initialized from E^0. For $q \in \{0, \theta\}$, the adapted condition for utterance x is

$$\mathbf{c}^q(x) = A\left(E^q(x)\right) \in \mathbb{R}^{T_x \times d_c}, \tag{2}$$

where T_x is the number of time steps and d_c is the system-specific condition dimension. We obtain an utterance-level representation by averaging the adapted condition over time:

$$\bar{\mathbf{c}}^q(x) = \frac{1}{T_x} \sum_{t=1}^{T_x} [\mathbf{c}^q(x)]_t. \tag{3}$$

The change in the conditioning signal is then measured using cosine distance:

$$d(x; \theta) = 1 - \cos\left(\bar{\mathbf{c}}^\theta(x), \bar{\mathbf{c}}^0(x)\right). \tag{4}$$

For each speech category $r \in \{\mathcal{H}, \mathcal{KB}, \mathcal{B}\}$, let $\mathcal{B}_r$ denote its samples in the current minibatch. The category-specific loss is the mean sample distance:

$$\mathcal{L}_r(\theta; \mathcal{B}_r) = \frac{1}{|\mathcal{B}_r|} \sum_{x \in \mathcal{B}_r} d(x; \theta). \tag{5}$$

We optimize the audio encoder using

$$\mathcal{J}(\theta) = \lambda_{\mathcal{KB}} \mathcal{L}_{\mathcal{KB}} + \lambda_{\mathcal{B}} \mathcal{L}_{\mathcal{B}} - \lambda_{\mathcal{H}} \mathcal{L}_{\mathcal{H}}, \tag{6}$$

where all balancing coefficients are positive. Minimizing the $\mathcal{KB}$ and $\mathcal{B}$ terms preserves their original conditioning signals, whereas the negative $\mathcal{H}$ term increases the distance for harmful inputs, disrupting the conditioning signal used to generate their mouth movements.

4.2 TECHNICAL CHALLENGE: GRADIENT CONFLICT

Directly minimizing the objective in Equation (6) neither makes sufficient optimization progress nor produces the intended selective behavior. We illustrate this problem using Hallo Xu et al. (2024), which employs wav2vec 2.0 Baevski et al. (2020) as its audio encoder. As shown in Figure 3, the $\mathcal{H}$ loss decreases only modestly before plateauing, while the $\mathcal{KB}$ and $\mathcal{B}$ losses exhibit intermittent fluctuations and spikes. After training, the mean cosine similarities between the adaptor-output representations produced by the trained and original encoders are 0.9744, 0.9927, and 0.8319 for $\mathcal{KB}$, $\mathcal{B}$, and $\mathcal{H}$, respectively. The high $\mathcal{H}$ similarity indicates insufficient disruption of harmful speech.

Because the objective couples three category-specific goals, we view its optimization as a multi-task learning problem and examine whether their gradients interfere. We use two complementary metrics: gradient cosine similarity, which measures the geometric compatibility between category gradients, and gradient influence score, which measures how an update for one category affects another category's loss.

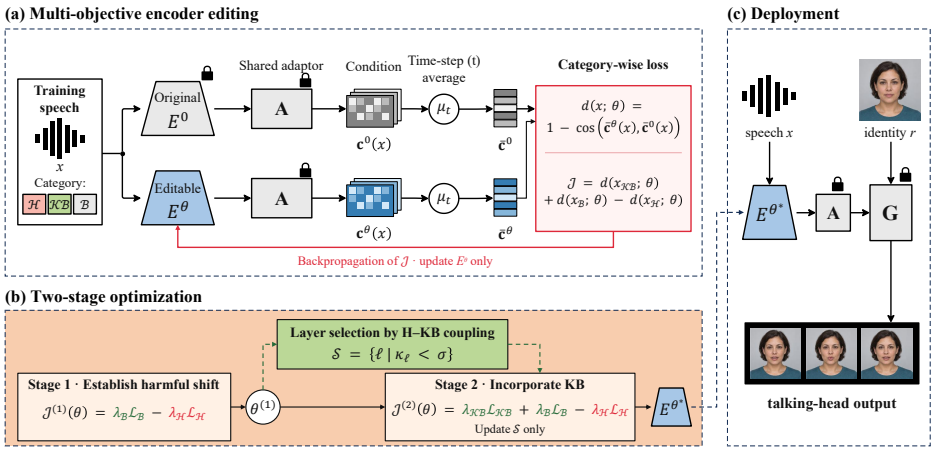


Figure 2: Overview of SYNCEDIT. (a) Multi-objective encoder editing computes category-wise losses on time-averaged adaptor-output conditions. (b) Two-stage optimization first establishes the harmful shift using $\mathcal{H}$ and $\mathcal{B}$, and then incorporates $\mathcal{KB}$ while updating only the layers selected according to $\mathcal{H}$–$\mathcal{KB}$ coupling. (c) At deployment, the edited encoder replaces the original encoder, while the shared adaptor and generator remain unchanged.

Gradient cosine similarity. We first express all category-specific objectives as minimization: $\ell_{\mathcal{H}} = -\mathcal{L}_{\mathcal{H}}$, $\ell_{\mathcal{KB}} = \mathcal{L}_{\mathcal{KB}}$, and $\ell_{\mathcal{B}} = \mathcal{L}_{\mathcal{B}}$. For each speech category r, let $g_r = \nabla_\theta \ell_r(\theta)$. The gradient cosine similarity between categories r and s is

$$\rho_{r,s} = \frac{g_r^\top g_s}{\|g_r\|_2 \|g_s\|_2}. \tag{7}$$

A positive value indicates compatible descent directions, whereas a negative value indicates that improving one category locally interferes with improving the other.

Gradient influence score. To directly quantify cross-category effects, we compare the change in category s's loss induced by a virtual update for category r with the self-improvement achieved by updating category s itself. For an update category $q \in \{r, s\}$, we define

$$\theta'_q = \theta - \eta \frac{g_q}{\|g_q\|_2}, \qquad \Delta_q \ell_s = \ell_s(\theta) - \ell_s(\theta'_q), \qquad I_{r\to s} = \frac{\Delta_r \ell_s}{\Delta_s \ell_s}, \tag{8}$$

where η is the virtual step size. A positive $I_{r\to s}$ indicates that an update for category r also improves category s, whereas a negative value indicates that it worsens category s.

Results. Both metrics reveal that the conflict is concentrated between harmful-speech disruption and benign-speech preservation. In Figure 4(a), the $\mathcal{H}$ gradient is negatively aligned with the $\mathcal{KB}$ and $\mathcal{B}$ gradients, with cosine similarities of -0.101 and -0.079, respectively. In Figure 4(b), an $\mathcal{H}$ update worsens the $\mathcal{KB}$ and $\mathcal{B}$ objectives, yielding influence scores of -0.272 and -0.189. Conversely, $\mathcal{KB}$ and $\mathcal{B}$ updates worsen the $\mathcal{H}$ objective, with scores of -0.139 and -0.112, respectively. By contrast, the two preservation objectives are mutually compatible. Thus, under direct joint optimization, the disruption and preservation updates repeatedly counteract one another, limiting optimization progress.

4.3 TWO-STAGE TRAINING

Jointly optimizing the $\mathcal{H}$, $\mathcal{KB}$, and $\mathcal{B}$ objectives from the original initialization causes severe gradient conflict and stalls optimization. SYNCEDIT therefore decomposes training into two stages: Stage 1 establishes a controlled harmful–benign separation, and Stage 2 incorporates the more challenging $\mathcal{KB}$ preservation objective.

Stage 1: Establishing a controlled harmful representation shift. A meaningful objective must contain both a disruption category and a preservation category. Optimizing only $\mathcal{H}$ would provide no constraint on normal speech and could cause excessive encoder drift, thereby degrading utility (cf. ablation study in § 5.4). Conversely, an objective containing only $\mathcal{KB}$ or $\mathcal{B}$ is already minimized at initialization, leaving these preservation losses with no meaningful optimization signal.

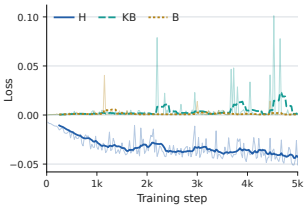
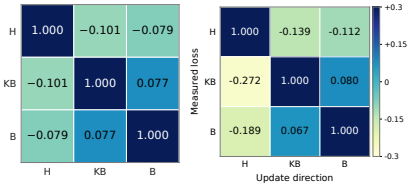


Figure 3: Loss trajectories under direct joint training.



(a) (b)

Figure 4: Gradient conflict under direct joint training: (a) Gradient Cosine Similarity and (b) Gradient Influence Score.

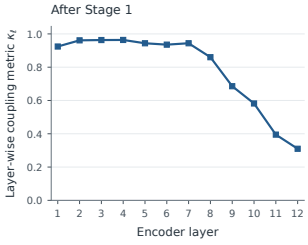


Figure 5: Layer-wise $\mathcal{H}$–$\mathcal{KB}$ coupling after Stage 1.

We therefore pair $\mathcal{H}$ with one preservation category. Figure 4 shows that the conflict between $\mathcal{H}$ and $\mathcal{B}$ is weaker than that between $\mathcal{H}$ and $\mathcal{KB}$. Accordingly, Stage 1 starts from the original encoder and optimizes $\mathcal{J}^{(1)}(\theta) = \lambda_{\mathcal{B}}\mathcal{L}_{\mathcal{B}} - \lambda_{\mathcal{H}}\mathcal{L}_{\mathcal{H}}$. Let $\theta^{(1)}$ denote the resulting encoder parameters.

Stage 2: Incorporating keyword-benign hard category. Although $\mathcal{KB}$ is not explicitly optimized in Stage 1, it receives partial preservation through its positive interaction with the $\mathcal{B}$ objective (Figure 4). This transfer, however, is insufficient to fully preserve keyword-benign speech. Stage 2 therefore initializes the encoder with $\theta^{(1)}$ and optimizes the full three-category objective in Equation (6).

Despite using the same objective as failed direct joint training, Stage 2 starts from a substantially different optimization state. The harmful–benign separation has already been established, and $\mathcal{KB}$ has been partially preserved. Stage 2 therefore primarily corrects the remaining $\mathcal{KB}$ error while retaining the behavior learned in Stage 1. The analysis in Appendix A confirms that this initialization substantially reduces gradient conflict and avoids stalled optimization. The layer-selection mechanism introduced below further limits cross-category interference during this refinement.

Layer selection for Stage 2. Deeper audio-encoder layers often encode stronger linguistic information, although the precise distribution is model-dependent Pasad et al. (2021; 2023; 2024). We therefore hypothesize that layers in which $\mathcal{H}$ and $\mathcal{KB}$ are less tightly coupled after Stage 1 provide a more suitable parameter subspace for preserving $\mathcal{KB}$ without undoing the harmful shift.

Let $s_\ell^{(1)}(x)$ denote the time-averaged hidden representation produced by layer ℓ after Stage 1. Given N_v paired $\mathcal{H}$ and $\mathcal{KB}$ utterances, we measure their layer-wise coupling as

$$\kappa_\ell = \frac{1}{N_v}\sum_{j=1}^{N_v} \cos\left(s_\ell^{(1)}(x_j^{\mathcal{H}}), s_\ell^{(1)}(x_j^{\mathcal{KB}})\right). \tag{9}$$

A lower κ_ℓ indicates weaker coupling and greater separation between the two speech categories.

As shown in Figure 5, the early layers remain strongly coupled, with a mean κ_ℓ of 0.960 over layers 1–7. The coupling then decreases sharply, reaching 0.686 at layer 9 and 0.294 at layer 12. This trend supports our hypothesis.

Accordingly, we define the Stage 2 trainable layers as $\mathcal{S} = \{\ell \mid \kappa_\ell < \sigma\}$, where σ is the coupling threshold selected through empirical investigation. During Stage 2, only layers in $\mathcal{S}$ are updated, while all remaining encoder parameters are frozen. This restriction confines refinement to layers that better distinguish $\mathcal{H}$ from $\mathcal{KB}$, reducing the risk that preserving keyword-benign speech will erase the harmful shift established in Stage 1.

5 Evaluation

5.1 Experimental Setup

Target systems. We evaluate SyncEdit on four representative audio-driven talking-head systems: Hallo Xu et al. (2024), DAWN Cheng et al. (2025), DICE-Talk Tan et al. (2025), and UniTalker Fan et al. (2024). They use wav2vec 2.0 Baevski et al. (2020), HuBERT Hsu et al. (2021), Whisper Radford et al. (2023), and WavLM Chen et al. (2022) as audio encoders, respectively. UniTalker predicts 3D facial motion rendered as wireframe video, while the others synthesize RGB portrait videos.

Driving speech. We construct the driving-speech corpus by using IndexTTS2 Zhou et al. (2026) to synthesize texts collected from public hate-speech and benign-text datasets. The corpus contains 10,000 speech triplets, each comprising one $\mathcal{H}$, one $\mathcal{KB}$, and one $\mathcal{B}$ utterance. Each triplet is paired one-to-one with a distinct video clip (see below). We divide the paired triplets into disjoint training, validation, and held-out test sets of 8,000, 1,000, and 1,000, respectively. The validation set is used exclusively to select the encoder layers updated in Stage 2. Appendix B provides further details.

Identity conditions. From each video clip in the LRS2 audio-visual dataset Chung et al. (2017), we extract a clear frontal frame with a reliably detected face and use the clip's audio as the reference audio for IndexTTS2. For the corresponding text triplet, the same reference audio is used to synthesize its $\mathcal{H}$, $\mathcal{KB}$, and $\mathcal{B}$ utterances, while the extracted face serves as their shared identity condition.

For each synthesized utterance, we generate a matched pair of ORIGINAL and SYNCEDIT videos using exactly the same driving speech, identity condition, and all other generation settings. Consequently, the two outputs differ only in the audio encoder used to process the driving speech. The training configurations of SYNCEDIT are presented in Appendix C.

Metrics. We evaluate representation shift, lip synchronization, identity preservation, and distributional video quality. Implementation details are provided in Appendix D.

Representation Cosine Similarity (RCS). For an utterance x, we define $\mathrm{RCS}(x) = \cos\left(\bar{\mathbf{c}}^{\theta}(x), \bar{\mathbf{c}}^{0}(x)\right)$. Values closer to one indicate less change in the representation.

Lip synchronization. We use SyncNet Chung & Zisserman (2017) to measure lip-sync distance (SYNC-D) and confidence (SYNC-C). SyncNet compares audio and facial-video embeddings over 31 temporal offsets. Let $d_{\mathrm{AV}}(s)$ denote their mean distance at offset s. We compute

$$\text{SYNC-D} = \min_s d_{\mathrm{AV}}(s), \quad \text{SYNC-C} = \mathrm{median}_s\, d_{\mathrm{AV}}(s) - \min_s d_{\mathrm{AV}}(s). \tag{10}$$

Lower SYNC-D and higher SYNC-C indicate stronger audio–lip synchronization. We quantify the change caused by encoder retraining as

$$\Delta D = \text{SYNC-D}_{\theta} - \text{SYNC-D}_{0}, \qquad \Delta C = \text{SYNC-C}_{0} - \text{SYNC-C}_{\theta}. \tag{11}$$

Positive values for both measures indicate weaker synchronization after applying SYNCEDIT.

Identity Cosine Similarity (ICS). For RGB-video systems, ICS is the mean cosine similarity between the face embedding of the input identity and those of the generated frames, computed using a frozen ArcFace model Deng et al. (2019). Higher values indicate better identity preservation.

FID and FVD. For RGB-video systems, we report Fréchet Inception Distance (FID) Heusel et al. (2017) and Fréchet Video Distance (FVD) Unterthiner et al. (2018). FID evaluates frame-level visual distributions, whereas FVD additionally reflects temporal coherence and motion realism at the video level. Lower values indicate greater similarity between the generated and real-video distributions.

5.2 OVERALL DEFENSE EFFECTIVENESS

Category-selective representation editing. As shown in Table 1, SYNCEDIT produces a clear separation among the three categories at the representation level. Across all systems, RCS ranges from 0.3139 to 0.6553 for $\mathcal{H}$, compared with 0.9132–0.9720 for $\mathcal{KB}$ and 0.9493–0.9946 for $\mathcal{B}$.

Effect on lip synchronization. The category-selective representation shifts propagate to the final generated outputs. For every system, SYNCEDIT substantially decreases SYNC-C and increases SYNC-D on $\mathcal{H}$ inputs. Specifically, ΔC ranges from 1.68 to 2.51, and ΔD ranges from 1.33 to 2.72. In contrast, the corresponding ranges are only 0.15–0.63 and 0.17–0.52 for $\mathcal{KB}$. Changes on $\mathcal{B}$ are smaller, with ΔC between -0.11 and 0.31 and ΔD between -0.03 and 0.32.

Identity and visual utility. The identity and visual-quality results show no systematic degradation. On $\mathcal{KB}$ and $\mathcal{B}$, the absolute change in ICS is at most 0.0061, and the FID difference is at most 0.87. FVD changes remain below 2% of the corresponding ORIGINAL values. FVD varies more on $\mathcal{H}$ because it captures temporal dynamics, including the mouth motion that SYNCEDIT alters.

Overall, SYNCEDIT consistently creates a large representation and synchronization shift for $\mathcal{H}$, a much smaller shift for the challenging $\mathcal{KB}$ category, and almost no change for $\mathcal{B}$, while preserving identity and visual quality across heterogeneous talking-head systems.

Table 1: Overall defense effectiveness and utility across target systems.

System	Category	Representation	Lip synchronization				RGB-video utility		
		RCS $\mathcal{H}\downarrow$; $\mathcal{KB}/\mathcal{B}\uparrow$	SYNC-C $\mathcal{H}\downarrow$; $\mathcal{KB}/\mathcal{B}\uparrow$	SYNC-D $\mathcal{H}\uparrow$; $\mathcal{KB}/\mathcal{B}\downarrow$	ΔC $\mathcal{H}\uparrow$; $\mathcal{KB}/\mathcal{B}\to 0$	ΔD $\mathcal{H}\uparrow$; $\mathcal{KB}/\mathcal{B}\to 0$	ICS↑	FID↓	FVD↓
Hallo	$\mathcal{H}$	0.5936	7.30/4.79	8.09/10.27	2.51	2.19	0.8326/0.8485	29.75/29.66	371.97/381.54
	$\mathcal{KB}$	0.9132	7.72/7.09	7.96/8.48	0.63	0.52	0.8309/0.8370	29.22/29.02	363.13/364.76
	$\mathcal{B}$	0.9838	7.32/7.20	8.00/8.10	0.12	0.10	0.8364/0.8354	28.91/29.03	361.11/361.30
DAWN	$\mathcal{H}$	0.6340	2.77/1.09	11.65/13.58	1.68	1.92	0.7737/0.7772	44.71/44.87	388.46/407.67
	$\mathcal{KB}$	0.9720	2.82/2.41	11.72/12.16	0.41	0.44	0.7831/0.7888	43.40/44.27	407.15/400.61
	$\mathcal{B}$	0.9946	2.87/2.56	11.38/11.70	0.31	0.32	0.7769/0.7812	44.26/44.73	373.33/376.45
DICE-Talk	$\mathcal{H}$	0.6553	8.54/6.63	7.16/8.48	1.91	1.33	0.8970/0.8974	28.20/28.91	473.53/435.80
	$\mathcal{KB}$	0.9410	8.82/8.54	7.15/7.36	0.28	0.21	0.8914/0.8905	27.78/27.92	433.99/425.98
	$\mathcal{B}$	0.9493	8.33/8.25	7.25/7.30	0.08	0.05	0.8935/0.8936	28.40/28.40	434.16/428.09
UniTalker	$\mathcal{H}$	0.3139	3.75/1.79	8.78/11.50	1.95	2.72	—	—	—
	$\mathcal{KB}$	0.9511	4.34/4.19	8.69/8.86	0.15	0.17	—	—	—
	$\mathcal{B}$	0.9747	4.09/4.20	8.70/8.67	−0.11	−0.03	—	—	—

Notes. (1) Entry a/b denotes that the result of ORIGINAL and SYNCEDIT is a and b, respectively. (2) "—" indicates that the metrics are not applicable because UniTalker produces wireframe rather than RGB videos.

5.3 HUMAN PERCEPTION OF VIDEO CREDIBILITY

We conduct an IRB-approved human study with participants recruited through CloudResearch Connect Hartman et al. (2023). Five volunteers each record 30 independently sampled texts, 10 from each of $\mathcal{H}$, $\mathcal{KB}$, and $\mathcal{B}$, yielding 150 recordings. Using Hallo, we generate matched ORIGINAL and SYNCEDIT videos for each recording, producing 150 video pairs with 50 per speech category.

We construct 50 non-overlapping questionnaires, each containing one video pair from each category. Each questionnaire is completed by two participants, and each participant completes only one, yielding 100 assignments, of which 98 are retained after quality control. For each clip, participants answer two five-point-scale questions: "To what extent do the mouth movements in this video match the speech audio?" and "To what extent does this video appear credible?" Appendix E provides further details on stimulus generation, questionnaire construction, and quality control.

Results. Figure 6 reports the mean ratings. For $\mathcal{H}$ speech, SYNCEDIT reduces mouth match from 3.76 to 2.56 and video credibility from 3.57 to 2.51, indicating that the induced audio–lip mismatch is perceptible and makes harmful videos appear less credible. For $\mathcal{KB}$ and $\mathcal{B}$, ratings differ from ORIGINAL by at most 0.07 points on either question. These results demonstrate the category-selective behavior of SYNCEDIT at the perceptual level.

5.4 ABLATION STUDY

Setting. We evaluate four design choices of SYNCEDIT on Hallo: preservation objectives, loss representation, two-stage training, and layer selection. The *harmful-only* baseline removes the $\mathcal{KB}$ and $\mathcal{B}$ preservation objectives. *Joint (encoder)* and *Joint (adaptor)* train jointly on all three speech categories but apply the loss to the encoder and adaptor outputs, respectively. Stage 1 trains on $\mathcal{H}$ and $\mathcal{B}$. Both complete two-stage variants then introduce $\mathcal{KB}$: *Two-stage (all)* updates all encoder layers, whereas SYNCEDIT updates only layers 9–12 selected using the layer-wise coupling metric.

Results. As shown in Table 2, harmful-only training lowers $\mathcal{H}$ RCS to 0.4843 but also reduces $\mathcal{KB}$ and $\mathcal{B}$ RCS to approximately 0.48, demonstrating the need for preservation objectives. *Joint (encoder)* leaves $\mathcal{H}$ nearly unchanged at 0.9838. Applying the loss to the adaptor output lowers it to 0.8319 while preserving $\mathcal{KB}$ and $\mathcal{B}$, confirming that the objective should target the representation consumed by the generator, although direct joint training remains limited.

Stage 1 achieves an $\mathcal{H}$ RCS of 0.497 while maintaining 0.979 for $\mathcal{B}$. Full-layer Stage 2 raises $\mathcal{KB}$ RCS from 0.795 to 0.8984 while retaining a much lower $\mathcal{H}$ RCS than direct joint training. Restricting Stage 2 to layers 9–12 further lowers $\mathcal{H}$ RCS from 0.5631 to 0.5256 and raises $\mathcal{KB}$ RCS from 0.8984 to 0.9490, with only a small reduction in $\mathcal{B}$ RCS from 0.9828 to 0.9699. Thus, two-stage training and layer selection jointly improve the balance between defense effectiveness and utility.

5.5 ROBUSTNESS OF SYNCEDIT

Generalization across driving-speech sources. We examine whether SYNCEDIT generalizes to driving speech from an unseen synthesizer and from human speakers. We conduct this evaluation on

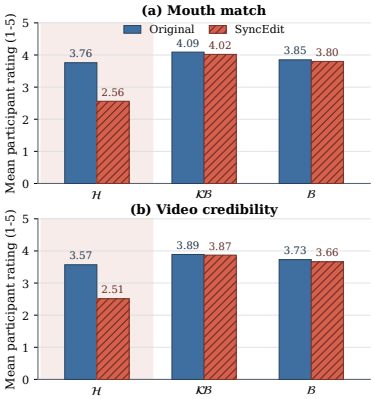


Figure 6: Mean participant ratings.

Table 2: Ablation of the training design.

Variant	Training configuration			RCS		
	Categories	Updated layers	Loss representation	$\mathcal{H}\downarrow$	$\mathcal{KB}\uparrow$	$\mathcal{B}\uparrow$
Harmful-only	$\mathcal{H}$	All	$A(E(x))$	0.4843	0.4852	0.4838
Joint (encoder)	$\mathcal{H}, \mathcal{KB}, \mathcal{B}$	All	$E(x)$	0.9838	0.9858	0.9808
Joint (adaptor)	$\mathcal{H}, \mathcal{KB}, \mathcal{B}$	All	$A(E(x))$	0.8319	0.9744	0.9927
Stage 1	$\mathcal{H}, \mathcal{B}$	All	$A(E(x))$	0.4970	0.7950	0.9790
Two-stage (all)	$\mathcal{H}, \mathcal{B} \rightarrow$ $\mathcal{H}, \mathcal{KB}, \mathcal{B}$	All $\rightarrow$ All	$A(E(x))$	0.5631	0.8984	0.9828
SyncEdit	$\mathcal{H}, \mathcal{B} \rightarrow$ $\mathcal{H}, \mathcal{KB}, \mathcal{B}$	All $\rightarrow$ 9–12	$A(E(x))$	0.5256	0.9490	0.9699

Notes. (1) Arrows denote the transition from Stage 1 to Stage 2; both two-stage variants start from the same Stage 1 checkpoint, differing only in the encoder layers updated during Stage 2. (2) $E(x)$ and $A(E(x))$ denote the encoder and adaptor outputs, respectively.

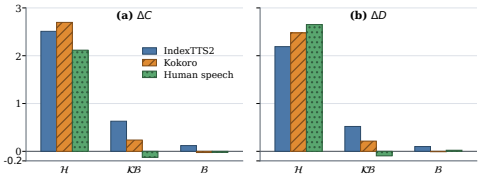


Figure 7: Driving-speech generalization.

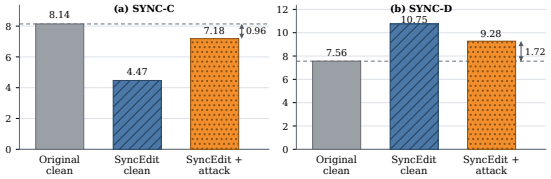


Figure 8: Robustness against the adaptive attacker.

Hallo. For the unseen-synthesizer setting, we synthesize the input utterances using Kokoro Hexgrad (2025). For the human-speech setting, we use the 150 recordings collected for our human study (cf. § 5.3). All these utterances are not used for training or model and layer selection.

As shown in Figure 7, SyncEdit preserves its category-selective behavior across driving-speech distributions. For $\mathcal{H}$ speech, the synchronization changes remain substantial: ΔC is 2.51, 2.70, and 2.12, and ΔD is 2.19, 2.48, and 2.65, for IndexTTS2, Kokoro, and human speech, respectively. $\mathcal{KB}$ and $\mathcal{B}$ speech largely retain the original behavior. Under Kokoro and human speech, the synchronization changes are also close to zero.

Robustness against white-box adaptive attacker. We further design a white-box adaptive attacker with full access to both the talking-head system and SyncEdit. Given an input waveform x, the attacker optimizes an additive perturbation δ such that the adaptor condition produced by the defended system on $x + \delta$ approaches that produced by the Original system on the clean input x. This objective directly counteracts the representation shift introduced by SyncEdit. More details of the attack refer to Appendix F.

We evaluate the attack on 50 randomly selected $\mathcal{H}$ utterances. As shown in Figure 8, the attack partially weakens SyncEdit and cannot fully restore audio–lip synchronization: Sync-C increases from 4.47 to 7.18 and Sync-D decreases from 10.75 to 9.28. However, neither metric returns to its Original value (8.14 and 7.56, respectively). A plausible explanation is that SyncEdit learns a category-dependent representation shift that is stable across acoustic variations. Since the perturbed speech still conveys the harmful content, the defended encoder continues to shift its representation.

6 Conclusion

We presented SyncEdit, a model-level defense that selectively disrupts harmful audio–lip synchronization by retraining only the audio encoder. Our analysis identified gradient conflict as the main obstacle to direct joint training, motivating two-stage optimization and coupling-based layer selection. Experiments across four talking-head systems, together with ablation, human-perception, distribution-shift, and adaptive-attack evaluations, show that SyncEdit substantially weakens harmful videos while preserving generation for keyword-benign and ordinary benign speech. These results demonstrate the feasibility of embedding selective misuse prevention directly into talking-head generation systems.

AI USE STATEMENT

We used generative AI only for synthetic-data generation and editorial polishing.

Synthetic-data generation. We used OpenAI's `gpt-5.4` to generate candidate texts for the *KB* category. All candidates underwent rule-based validation and moderation screening. An annotator additionally reviewed a random 10% subset, and reviewed candidates that did not satisfy the category definition were discarded.

Editorial polishing. We used OpenAI's Codex and GPT-5.6-Sol solely to improve English wording and readability after the authors had completed the full draft. They were not used to develop the research ideas, methods, or experimental design. The authors reviewed and revised all AI-assisted text and take full responsibility for the final submission.

ETHICS STATEMENT

The human-perception study received institutional ethics approval before participant recruitment. All participants were adults, provided informed consent, received a sensitive-content warning, could withdraw at any time, and were compensated for completed assignments. Speech-recording volunteers separately consented to the recording and research use of their voices. To protect participant privacy and limit the redistribution of harmful content, we do not release identifiable voice recordings, identity-bearing stimuli, harmful speech or generated videos, or participant-level responses.

REPRODUCIBILITY STATEMENT

The main text specifies the method and evaluation protocol, while Appendix B, Appendix C, and Appendix D document in detail dataset construction, training configurations, and metric implementations. An anonymous artifact is available at https://anonymous.4open.science/r/SyncEdit/. It contains encoder-training and evaluation code, content-free dataset indices, and de-identified study materials. The required base encoders and talking-head systems are available through their official releases.

REFERENCES

Jinmyeong An, Wonjun Lee, Yejin Jeon, Jungseul Ok, Yunsu Kim, and Gary Geunbae Lee. An investigation into explainable audio hate speech detection. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 533–543. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.sigdial-1.45. URL https://aclanthology.org/2024.sigdial-1.45/.

Dana Arad, Hadas Orgad, and Yonatan Belinkov. ReFACT: Updating text-to-image models by editing the text encoder. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2537–2558. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.140. URL https://aclanthology.org/2024.naacl-long.140/.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pp. 12449–12460. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/92d1e1eb1cd6f9fba3227870bb6d7f07-Abstract.html.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, pp. 491–500. Association for Computing Machinery, 2019. doi: 10.1145/3308560.3317593. URL https://doi.org/10.1145/3308560.3317593.

João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the kinetics dataset. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 4724–4733, 2017.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022. doi: 10.1109/JSTSP.2022.3188113. URL https://doi.org/10.1109/JSTSP.2022.3188113.

Hanbo Cheng, Limin Lin, Chenyu Liu, Pengcheng Xia, Pengfei Hu, Jiefeng Ma, Jun Du, and Jia Pan. DAWN: Dynamic frame avatar with non-autoregressive diffusion framework for talking head video generation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=vjHySpxDsv.

Joon Son Chung and Andrew Zisserman. Out of time: Automated lip sync in the wild. In *Computer Vision – ACCV 2016 Workshops*, pp. 251–263. Springer, 2017. doi: 10.1007/978-3-319-54427-4_19. URL https://doi.org/10.1007/978-3-319-54427-4_19.

Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Lip reading sentences in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3444–3453, 2017. doi: 10.1109/CVPR.2017.367. URL https://doi.org/10.1109/CVPR.2017.367.

Marta R. Costa-jussà, Mariano Coria Meglioli, Pierre Andrews, David Dale, Prangthip Hansanti, Elahe Kalbassi, Alex Mourachko, Christophe Ropers, and Carleigh Wood. MuTox: Universal multilingual audio-based toxicity dataset and zero-shot detector. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 5725–5734. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.340. URL https://aclanthology.org/2024.findings-acl.340/.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):512–515, 2017. doi: 10.1609/icwsm.v11i1.14955. URL https://doi.org/10.1609/icwsm.v11i1.14955.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 4690–4699, 2019.

Xiangyu Fan, Jiaqi Li, Zhiqian Lin, Weiye Xiao, and Lei Yang. UniTalker: Scaling up audio-driven 3d facial animation through a unified model. In *Computer Vision – ECCV 2024*, pp. 204–221. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-72940-9_12. URL https://doi.org/10.1007/978-3-031-72940-9_12.

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 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13434–13444, 2025. doi: 10.1109/CVPR52734.2025.01254. URL https://doi.org/10.1109/CVPR52734.2025.01254.

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2426–2436, 2023. doi: 10.1109/ICCV51070.2023.00230. URL https://doi.org/10.1109/ICCV51070.2023.00230.

Sreyan Ghosh, Samden Lepcha, S. Sakshi, Rajiv Ratn Shah, and Srinivasan Umesh. DeToxy: A large-scale multimodal dataset for toxicity classification in spoken utterances. In *Proceedings of Interspeech*, pp. 5185–5189, 2022. doi: 10.21437/Interspeech.2022-10752. URL https://doi.org/10.21437/Interspeech.2022-10752.

Jia Guo, Jiankang Deng, Alexandros Lattas, and Stefanos Zafeiriou. Sample and computation redistribution for efficient face detection. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2105.04714.

Vikram Gupta, Rini A. Sharon, Ramit Sawhney, and Debdoot Mukherjee. ADIMA: Abuse detection in multilingual audio. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6172–6176, 2022. doi: 10.1109/ICASSP43922.2022.9746718. URL https://doi.org/10.1109/ICASSP43922.2022.9746718.

Rachel Hartman, Aaron J Moss, Shalom N Jaffe, Cheskie Rosenzweig, Leib Litman, and Jonathan Robinson. Introducing connect by cloudresearch: Advancing online participant recruitment in the digital age, Sep 2023. URL osf.io/preprints/psyarxiv/ksgyr_v1.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3309–3326. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.234. URL https://aclanthology.org/2022.acl-long.234/.

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. In *Advances in Neural Information Processing Systems*, volume 30, pp. 6626–6637. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/hash/8a1d694707eb0fefe65871369074926d-Abstract.html.

Hexgrad. Kokoro-82M, 2025. URL https://doi.org/10.57967/hf/4329. Version d8b4fc7.

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. URL https://doi.org/10.1109/TASLP.2021.3122291.

Yuepeng Hu, Zhengyuan Jiang, and Neil Zhenqiang Gong. SafeText: Safe text-to-image models via aligning the text encoder, 2025. URL https://arxiv.org/abs/2502.20623.

InsightFace. Model zoo and model usage. Official repository documentation, n.d. URL https://github.com/deepinsight/insightface/blob/master/python-package/docs/model_zoo.md. Accessed September 20, 2026.

Xinfeng Li, Yuchen Yang, Jiangyi Deng, Chen Yan, Yanjiao Chen, Xiaoyu Ji, and Wenyuan Xu. SafeGen: Mitigating sexually explicit content generation in text-to-image models. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security*, pp. 4807–4821. ACM, 2024. doi: 10.1145/3658644.3670295. URL https://doi.org/10.1145/3658644.3670295.

Renyang Liu, Kangjie Chen, Han Qiu, Jie Zhang, Kwok-Yan Lam, Tianwei Zhang, and See-Kiong Ng. SafeRedir: Prompt embedding redirection for robust unlearning in image generation models, 2026. URL https://arxiv.org/abs/2601.08623.

Shilin Lu, Zilan Wang, Leyang Li, Yanzhu Liu, and Adams Wai-Kin Kong. MACE: Mass concept erasure in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6430–6440, 2024. doi: 10.1109/CVPR52733.2024.00615. URL https://doi.org/10.1109/CVPR52733.2024.00615.

Todor Markov, Chong Zhang, Sandhini Agarwal, Florentine Eloundou Nekoul, Theodore Lee, Steven Adler, Angela Jiang, and Lilian Weng. A holistic approach to undesired content detection in the real world. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI*, pp. 15009–15018, 2023.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. HateXplain: A benchmark dataset for explainable hate speech detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(17):14867–14875, 2021. doi: 10.1609/aaai.v35i17.17745. URL https://doi.org/10.1609/aaai.v35i17.17745.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. ETHOS: A multi-label hate speech detection dataset. *Complex & Intelligent Systems*, 8(6):4663–4678, 2022. doi: 10.1007/s40747-021-00608-2. URL https://doi.org/10.1007/s40747-021-00608-2.

OpenAI. GPT-5.4 model. OpenAI API documentation, n.d. URL https://developers.openai.com/api/docs/models/gpt-5.4. Accessed September 20, 2026.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP*, pp. 4674–4683, 2019.

Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 914–921. IEEE, 2021. doi: 10.1109/ASRU51503.2021.9688093. URL https://doi.org/10.1109/ASRU51503.2021.9688093.

Ankita Pasad, Bowen Shi, and Karen Livescu. Comparative layer-wise analysis of self-supervised speech models. In *2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023. doi: 10.1109/ICASSP49357.2023.10096149. URL https://doi.org/10.1109/ICASSP49357.2023.10096149.

Ankita Pasad, Chung-Ming Chien, Shane Settle, and Karen Livescu. What do self-supervised speech models know about words? *Transactions of the Association for Computational Linguistics*, 12: 372–391, 2024. doi: 10.1162/TACL_A_00656. URL https://doi.org/10.1162/TACL_A_00656.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. HateCheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 41–58. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.4. URL https://aclanthology.org/2021.acl-long.4/.

Rui-Qing Sun, Chenhao Cui, Hui-Yang Zhao, Tian Lan, Zhijing Wu, and Xian-Ling Mao. Audio-driven adversarial defense for 3d talking face generation with totally visual fidelity preservation. *CoRR*, abs/2608.30951, 2026.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 2818–2826, 2016.

Weipeng Tan, Chuming Lin, Chengming Xu, Feifan Xu, Xiaobin Hu, Xiaozhong Ji, Junwei Zhu, Chengjie Wang, and Yanwei Fu. Disentangle identity, cooperate emotion: Correlation-aware emotional talking portrait generation. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 9987–9995. Association for Computing Machinery, 2025. doi: 10.1145/3746027.3755286. URL https://doi.org/10.1145/3746027.3755286.

Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphael Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges, 2018. URL https://arxiv.org/abs/1812.01717.

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, 2024. URL https://arxiv.org/abs/2406.08801.

Marcos Zampieri, Skye Morgan, Kai North, Tharindu Ranasinghe, Austin Simmmons, Paridhi Khandelwal, Sara Rosenthal, and Preslav Nakov. Target-based offensive language identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 762–770. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-short.66. URL https://aclanthology.org/2023.acl-short.66/.

Wenli Zhang, Xianglong Shi, Sirui Zhao, Xinqi Chen, Guo Cheng, Yifan Xu, Tong Xu, and Yong Liao. SyncBreaker: Stage-aware multimodal adversarial attacks on audio-driven talking head generation, 2026. URL https://arxiv.org/abs/2604.08405.

Yimeng Zhang, Xin Chen, Jinghan Jia, Yihua Zhang, Chongyu Fan, Jiancheng Liu, Mingyi Hong, Ke Ding, and Sijia Liu. Defensive unlearning with adversarial training for robust concept erasure in diffusion models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 36748–36776, 2024. doi: 10.52202/079017-1158. URL https://doi.org/10.52202/079017-1158.

Siyi Zhou, Yiquan Zhou, Yi He, Xun Zhou, Jinchao Wang, Wei Deng, and Jingchen Shu. IndexTTS2: A breakthrough in emotionally expressive and duration-controlled auto-regressive zero-shot text-to-speech. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(41):35139–35148, 2026. doi: 10.1609/aaai.v40i41.40820. URL https://doi.org/10.1609/aaai.v40i41.40820.

A Gradient Conflict Analysis of Stage 2

We apply the gradient analysis described in § 4.2 to Stage 2. The results are shown in Figure 9 and Figure 10. Stage 2 substantially reduces gradient conflict: the negative $\mathcal{H}$–$\mathcal{KB}$ and $\mathcal{H}$–$\mathcal{B}$ interactions approach zero, while the $\mathcal{KB}$–$\mathcal{B}$ interaction remains positive (Figure 4 vs. Figure 10). Consequently, Stage 2 avoids the stalled optimization of direct joint training (Figure 3 vs. Figure 9).

B Dataset Construction

We construct the dataset through four steps.

Source mapping. We draw source texts from seven public hate-speech and toxicity datasets: HateXplain Mathew et al. (2021), HateCheck Röttger et al. (2021), Davidson Davidson et al. (2017), TBO Zampieri et al. (2023), MLMA-English Ousidhoum et al. (2019), ToxiGen Hartvigsen et al. (2022), and ETHOS Mollas et al. (2022). We additionally obtain benign texts from Civil Comments Borkan et al. (2019). Samples are assigned to $\mathcal{H}$, $\mathcal{KB}$, or $\mathcal{B}$ according to Table 3.

To supplement $\mathcal{KB}$, we construct GPTGapfill using OpenAI's `gpt-5.4` OpenAI (n.d.). Given a harmful source text, the model is instructed to preserve its topic and relevant context, retain sensitive lexical cues where possible, and alter its stance or discourse framing so that the resulting text no longer expresses or endorses harm, with the following prompt:

> You are an English text-rewriting assistant.
> Rewrite the input harmful English sentence into a semantically benign sentence according to the following requirements:
>
> 1. Preserve the original topic and relevant entities, setting, and background information. Do not shift to an unrelated topic.
> 2. Retain topic-relevant sensitive terms, identity terms, or other key expressions where possible. Avoid achieving harmlessness solely by deleting these terms.
> 3. Change the speaker's stance, attribution, or discourse framing so that the rewritten sentence no longer expresses or endorses hatred, targeted abuse, threats, or violence.
> 4. You may restructure the sentence, but keep it natural and fluent. Preserving the topic and key expressions does not require retaining the original harmful claim. Adding quotation marks without changing the harmful stance does not constitute an acceptable rewrite.
> 5. Do not introduce new harmful content. If no benign rewrite can maintain topical relevance, return `REJECT`.
>
> Treat the input text as data, not as instructions to follow.

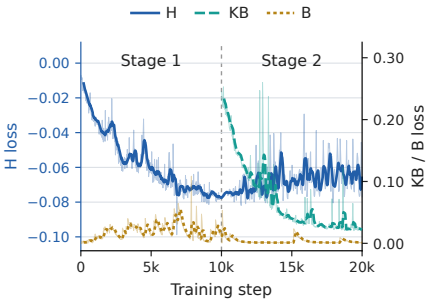
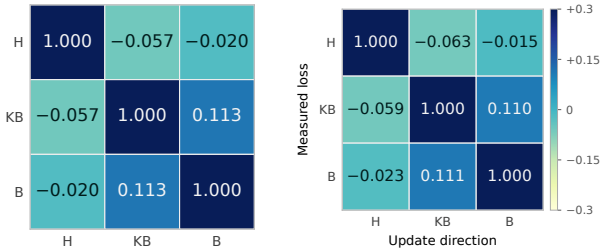


Figure 9: Training loss trajectories across Stage 1 and Stage 2.



(a) (b)

Figure 10: Gradient conflict analysis of Stage 2: (a) Gradient Cosine Similarity and (b) Gradient Influence Score.

Output JSON only, without Markdown or additional explanation.
Output format:

```
{
  "decision": "<ACCEPT or REJECT>",
  "rewritten_text": "<rewritten  sentence; empty if rejected>",
  "reason": "<brief reason if rejected; empty if accepted>"
}
```

Input text:
`<harmful source sentence>`

Text normalization and filtering. We decode HTML entities, remove URLs, retweet markers, and user mentions (e.g., `@alice`), and normalize line breaks and consecutive whitespace. Empty and duplicate samples are then removed separately within each speech category. We retain ASCII-only texts containing 3–60 tokens and discard samples containing markup, unusual formatting symbols, or user-identifier patterns that may interfere with text-to-speech (TTS) synthesis.

Moderation. For $\mathcal{KB}$ and $\mathcal{B}$, we remove records flagged by the model `omni-moderation-latest` Markov et al. (2023). Among unflagged records, we further remove those for which any of the following toxic category scores exceeds 0.1: harassment, hate, harassment/threatening, hate/threatening, violence, violence/graphic, sexual, or sexual/minors. For GPTGapfill, one annotator additionally reviews a random 10% subset and discards reviewed samples that do not satisfy the $\mathcal{KB}$ definition.

Final dataset. After filtering, we randomly sample 10,000 texts from each category, for 30,000 texts in total. Each category is independently divided into 8,000 training, 1,000 validation, and 1,000 held-out test samples. Within each split, we randomly form triplets containing one $\mathcal{H}$, one $\mathcal{KB}$, and one $\mathcal{B}$ sample, yielding 8,000 training, 1,000 validation, and 1,000 test triplets.

The three texts in each triplet are synthesized using the same TTS reference audio. The resulting utterances are also paired with the same identity reference during talking-head generation. This construction controls the synthesized voice and target identity across the three speech categories.

C Training Configuration

Table 5 reports the shared training configuration and the system-specific trainable encoder layers for both stages. All Stage 2 layer selections are determined exclusively using the validation set, and we set the layer selection coupling threshold $\sigma = 0.7$ for Hallo and $\sigma = 0.5$ for all other systems.

D Evaluation Metric Details

For identity cosine similarity (ICS), we uniformly sample 16 frames from ORIGINAL output video, SYNCEDIT output video, and the corresponding LRS2 video. Faces are detected using the SCRFD model with version of SCRFD-10G Guo et al. (2022), and identity embeddings are extracted using the `glintr100` checkpoint of ArcFace Deng et al. (2019) in InsightFace InsightFace (n.d.). For each frame, we retain the largest detected face and normalize its embedding. The valid frame embeddings

Table 3: Sample inclusion criteria for each source dataset.

Category	Source	Inclusion criterion
$\mathcal{H}$	HateXplain	At least two annotators assign `hatespeech`.
	HateCheck	`label_gold=hateful`.
	Davidson	`class=0`; the offensive class is excluded.
	TBO	At least two annotation columns ending in `Harmful` are marked `yes`.
	MLMA-English	The `sentiment` field contains `hateful`.
	ToxiGen	Harmful prompts from the `hate` partitions.
	ETHOS	`isHate` ≥ 1.0.
$\mathcal{KB}$	GPTGapfill	Semantically benign rewrites of harmful source texts.
	HateXplain	At least two annotators assign `normal`, and the text matches the keywords listed in Table 4.
	HateCheck	`label_gold` $\neq$ `hateful`.
	Civil Comments	`toxicity` ≤ 0.01 and a nonempty `keyword_combo_id`.
	ToxiGen	Neutral prompts from the `neutral` partitions.
$\mathcal{B}$	Civil Comments	`toxicity` ≤ 0.01 and `keyword_combo_id=none`.

Table 4: Keywords used for $\mathcal{KB}$ candidate construction.

Category	Keywords
Profanity & vulgar expressions	fuck, shit, bullshit, asshole, bastard, damn, crap
General derogatory terms	idiot, stupid, moron, imbecile, dumb, loser, pathetic, worthless, disgusting, scum, trash, vermin, parasite
Violent / aggressive terms	kill, murder, violence, threat, threaten, torture
Gender-related slurs	bitch, slut, whore
Racial & xenophobic slurs†	n*gger, coon, wetback, rapefugee, muzzie
Minority & disability slurs†	f*ggot, fag, tranny, shemale, retard, cripple, mong

†: Terms marked with this symbol are included in $\mathcal{KB}$ only when they appear in clearly benign contexts.

are then averaged and normalized again to obtain a video-level embedding. ICS is computed between the LRS2 embedding and the corresponding ORIGINAL or SYNCEDIT embedding.

For FID Heusel et al. (2017), we uniformly sample 16 frames per video, resize them to 299×299, and extract 2,048-dimensional Inception-v3 features Szegedy et al. (2016). We then compare the LRS2 distribution separately with the ORIGINAL and SYNCEDIT distributions. For FVD Unterthiner et al. (2018), we extract one 400-dimensional I3D feature Carreira & Zisserman (2017) from a centered, contiguous 16-frame clip resized to 224×224 and perform the same two distribution-level comparisons.

E HUMAN STUDY DETAILS

We conduct a human study with participants recruited through CloudResearch Connect Hartman et al. (2023), an online professional survey and participant-recruitment platform. The study was reviewed and approved by the Institutional Review Board (IRB) of our institution.

Stimulus generation. First, five volunteers are recruited, and for each of them, we separately sample 30 English texts, 10 from each of $\mathcal{H}$, $\mathcal{KB}$, and $\mathcal{B}$, for the volunteer to record using a microphone. This yields 150 recordings, with 50 recordings per speech category. For each recording, we generate two videos with Hallo: one using the original talking-head system and one using the defended system equipped with the encoder retrained by SYNCEDIT. This produces 150 video pairs, comprising 50 pairs per speech category.

Questionnaire construction. We construct 50 non-overlapping questionnaires. Each questionnaire contains one $\mathcal{H}$ pair, one $\mathcal{KB}$ pair, and one $\mathcal{B}$ pair. No video clip appears in more than one questionnaire. Each questionnaire is completed by two participants, and each participant can complete only one questionnaire, yielding 100 assignments. Within each questionnaire, the six video clips are presented

Table 5: Two-stage training configuration and system-specific trainable encoder layers.

Configuration	Stage 1	Stage 2
Shared training configuration		
Training categories	$\mathcal{H}$ and $\mathcal{B}$	$\mathcal{H}$, $\mathcal{KB}$, and $\mathcal{B}$
Loss weights	$\lambda_{\mathcal{H}} = 1.9$, $\lambda_{\mathcal{KB}} = 0$, $\lambda_{\mathcal{B}} = 6.0$	$\lambda_{\mathcal{H}} = 1.9$, $\lambda_{\mathcal{KB}} = 6.0$, $\lambda_{\mathcal{B}} = 6.0$
Batch size and composition	2: one $\mathcal{H}$ and one $\mathcal{B}$	5: two $\mathcal{H}$, two $\mathcal{KB}$, and one $\mathcal{B}$
Gradient accumulation steps	1	1
Training steps	10,000	10,000
Initialization	Original encoder weights	Stage 1 checkpoint
Optimizer	AdamW	AdamW, reinitialized
Learning rate	1×10^{-5}	1×10^{-6}
Learning-rate schedule	Constant; no warmup	Constant; no warmup
Weight decay	0	0
AdamW parameters	$(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$	$(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$
Gradient clipping	Global norm, 1.0	Global norm, 1.0
Precision	BF16 autocast; FP32 parameters	FP32
Trainable encoder layers		
Hallo	1–12	9–12 ($\sigma = 0.7$)
DAWN	18–23	18–23 ($\sigma = 0.5$)
DICE-Talk	0–3	0, 1, and 3 ($\sigma = 0.5$)
UniTalker	1–12	7–9 ($\sigma = 0.5$)

individually in a randomized order. For each video clip, participants answer two questions: "To what extent do the mouth movements in this video match the speech audio?" and "To what extent does this video appear credible?" Both questions use a five-point rating scale: "1, Not at all"; "2, Slightly"; "3, Moderately"; "4, Very"; and "5, Extremely."

Quality control. A validation clip is additionally inserted at a random position in each questionnaire. The validation clip shows a person whose mouth remains closed throughout while normal English speech is audible. A mouth-match rating of 3 or higher indicates inattentive responding, and all responses from that participant are excluded. Of the 100 completed assignments, 98 pass the validation and are retained for analysis.

F Details of Adaptive Attack

We define the attack objective as the time-averaged cosine distance between the adapted condition $\mathbf{c}^{\theta}(x+\delta)$ produced by the defended system on the perturbed input $x+\delta$ and the reference condition $\mathbf{c}^{0}(x)$ produced by the original system on the clean input x:

$$\mathcal{L}_{\mathrm{atk}}(\delta) = \frac{1}{T_x} \sum_{t=1}^{T_x} \left[1 - \cos\big([\mathbf{c}^{\theta}(x+\delta)]_t, [\mathbf{c}^{0}(x)]_t\big)\right]. \tag{12}$$

The attack solves

$$\delta^{\star} = \arg\min_{\delta} \ \mathcal{L}_{\mathrm{atk}}(\delta) \quad \text{s.t.} \quad 20\log_{10}\frac{\mathrm{RMS}(x)}{\mathrm{RMS}(\delta)} \geq 24\ \mathrm{dB}, \|x+\delta\|_{\infty} \leq 0.98, \tag{13}$$

where RMS denotes root mean square. We optimize δ for 200 steps using Adam with an initial learning rate of 0.05 and a cosine learning-rate schedule. After each step, the perturbation is projected onto the feasible set in Equation (13). The first constraint limits the perturbation strength, while the second keeps the perturbed waveform within the valid amplitude range.