ESIAN: TRAINING-FREE CONTINUAL SPEAKER IDENTITY ERASURE FOR ZERO-SHOT TTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Zero-shot text-to-speech (ZS-TTS) models can clone an arbitrary voice from only a few seconds of reference audio, even when the speaker was absent from the training data. When a speaker requests that their voice no longer be cloned, the provider therefore needs to address this request regardless of whether that speaker's recordings were used for training. We study erasing target-speaker identity during generation while preserving voice cloning for other speakers. Finetuning-based unlearning approaches update the model's weights for every removal request, a cost that becomes impractical as requests accumulate. Inference-time alternatives avoid weight updates but assume that the incoming prompt is already known to belong to a target speaker. As new requests continue to arrive, maintaining erasure for previously enrolled speakers poses an additional challenge. We propose ESIAN (Erasing Speaker Identity via Adaptive gatiNg), a training-free framework that treats an erasure request as a single enrollment step. From several minutes of enrollment recordings, ESIAN derives a closed-form eraser together with an intrinsic gate that determines, for every incoming prompt, how strongly the eraser should be applied. Our framework enables continual speaker identity erasure without updating the backbone weights or revisiting the utterances of previous requests. Experiments demonstrate that ESIAN achieves erasure comparable to finetuning-based unlearning while preserving voice-cloning quality for the remaining speakers, and maintains low average target-speaker similarity as the enrolled set grows to a hundred speakers.

1 INTRODUCTION

The right to be forgotten and its recognition in the General Data Protection Regulation (GDPR) highlight the importance of individuals' control over their personal data (Mantelero, 2013; Voigt & von dem Bussche, 2017). Speech generation presents a particular challenge to this control. Zero-shot text-to-speech (ZS-TTS) models clone an arbitrary voice from only a few seconds of reference audio, called a *prompt*, so a voice can be replicated even if its speaker never appeared in the training data (Le et al., 2023; Wang et al., 2023). Simply excluding someone from training therefore does not prevent their voice from being cloned. When a speaker requests that their voice no longer be cloned, the provider needs to suppress reproduction of that voice in the deployed system while preserving voice cloning for other speakers. As the erasure requests arrive sequentially, the provider needs to suppress the voices of an expanding set of speakers (Figure 1).

Existing approaches to suppressing target-speaker identity fall into two categories. Finetuning-based unlearning updates the model so that target-speaker prompts no longer yield their voice (Kim et al., 2025; Lertpetchpun et al., 2026). Repeating this training for every removal request incurs substantial computational cost. Inference-time alternatives, such as TruS (Lee et al., 2026b), leave the weights untouched and suppress the target identity during generation. TruS, however, assumes that the incoming prompt is already known to belong to a target speaker, leaving prompt identification to be handled separately at deployment.

To address these challenges, we propose ESIAN (Erasing Speaker Identity via Adaptive gatiNg), a training-free framework for continual speaker identity erasure that combines a closed-form projection with an intrinsic gate.

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.

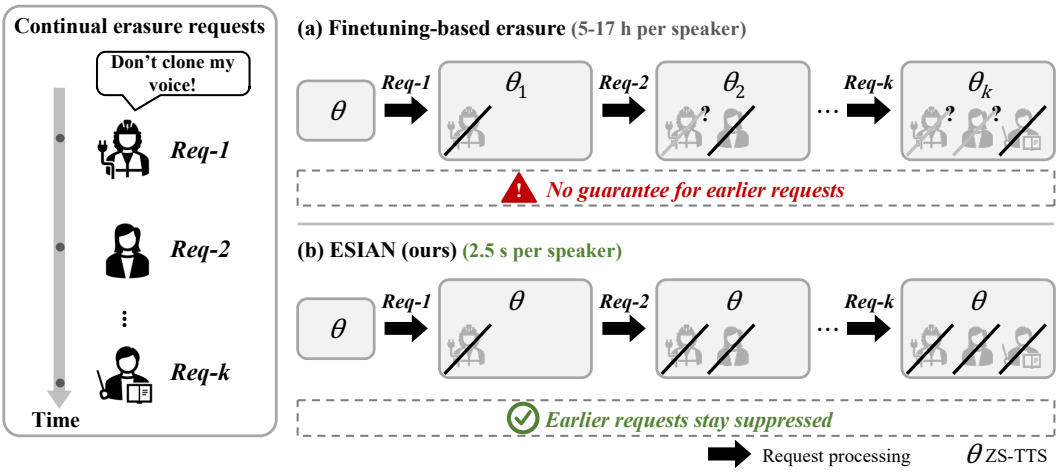

Figure 1: Erasure requests arrive one after another. (a) Finetuning-based erasure (Kim et al., 2025) updates the model weights at every request, θ_1 through θ_k. (b) ESIAN leaves the weights unchanged and incorporates new requests by accumulating speaker statistics. Crossed-out speakers indicate suppressed identities; question marks indicate uncertain suppression of earlier targets. Suppression of earlier requests is assessed empirically through request-group averages and individual-speaker tracking. The approximate 2.5 s cost denotes single-speaker enrollment compute, excluding I/O and gate refitting (Appendix B).

In ZS-TTS models where the reference audio is injected through a single linear projection, a speaker's identity reaches the backbone along that conditioning path. ESIAN constructs a closed-form projection from the dominant directions extracted from recordings provided with erasure requests and applies it during generation without retraining the backbone. In the continual setting, ESIAN incorporates new erasure requests by accumulating per-speaker statistics, without revisiting recordings from previous requests.

Applying the projection to every prompt, however, also suppresses remaining speakers because their conditioning features overlap with those of target speakers. ESIAN therefore compares each prompt's internal activation signature with representative signatures of target speakers to determine its erasure strength. Since remaining-speaker recordings are unavailable, generations produced without a prompt serve as synthetic negatives for score normalization and threshold calibration.

The main contributions are as follows:

- We propose ESIAN, a training-free framework for erasing speaker identity in ZS-TTS that adaptively modulates erasure strength for each generation request using a closed-form projection. We design a gate that estimates the appropriate erasure strength from internal activations, thus removing reliance on external detection models during inference.
- ESIAN handles sequential erasure requests with no weight update and no oracle telling it which prompts to erase. Requests are incorporated without revisiting earlier audio, and we establish conditions under which the resulting eraser is identical to one recomputed on the union of all requests.
- We demonstrate erasure comparable to finetuning-based unlearning while preserving voice-cloning quality for the remaining speakers, and show that ESIAN maintains low average target-speaker similarity as the target set grows to a hundred speakers.

2 Related Work

Machine Unlearning Machine unlearning removes the influence of designated data while preserving remaining utility (Bourtoule et al., 2021; Nguyen et al., 2025). Retraining with the data excluded is costly, motivating approximate weight-update methods (Golatkar et al., 2020; Thudi

et al., 2022). In generative models, related work studies erasing concepts and styles from diffusion models (Gandikota et al., 2023; Han et al., 2025). Continual unlearning further addresses sequential requests, where repeated updates risk utility degradation (Shi et al., 2025) and the reappearance of removed content (Lee et al., 2026a; Thakral et al., 2025; Xu et al., 2026). O^3 combines continual unlearning with an inference-time detector that controls the activation and strength of an unlearning LoRA (Gao et al., 2025). Whereas O^3 trains the adapters and detector, ESIAN constructs a closed-form eraser and an intrinsic gate for ZS-TTS without such training.

In speech generation, Teacher-Guided Unlearning (TGU) and Sample-Guided Unlearning (SGU) (Kim et al., 2025) formulate speaker identity unlearning for ZS-TTS, finetuning the model so that forget-speaker prompts yield random voices, at substantial cost per request. SGSP (Lertpetchpun et al., 2026) extends this line to StyleTTS2 with encoder-guided and contrastive objectives, and reports that erasure degrades at a hundred speakers as identities overlap. Inference-time alternatives leave the weights untouched: TruS (Lee et al., 2026b) steers hidden activations during every generation but presumes the prompt is already known to belong to a target speaker, and a baseline filters prompts through an external speaker-verification system (Lertpetchpun et al., 2026). In either case, detection is outsourced or assumed.

Feature Space Erasure Feature-space erasure projects representations onto the orthogonal complement of concept directions without updating model weights. In NLP, this strips protected attributes from representations by nullspace projection, by an adversarial objective, or in closed form (Ravfogel et al., 2020; 2022; Belrose et al., 2023); the same idea supports gradient-free class forgetting in classifiers (Kodge et al., 2024). Closest to our approach, ESC (Lee et al., 2025) applies this to unlearning: the principal directions of the forget-class features, obtained by SVD, are removed by a projection applied uniformly to every input. These methods work where the forget and remaining concepts occupy separable directions, as in classification. In the conditioning features studied here, target and remaining speakers are almost entirely entangled, so a uniform projection erases the remaining speakers along with the target speakers. We therefore gate the projection per prompt, calibrated on the model's own generations.

3 PRELIMINARIES

Background VoiceBox (Le et al., 2023) is a non-autoregressive speech infilling model trained with conditional flow matching (Lipman et al., 2023). Its Transformer predicts a vector field $v_t(x_t, x_{\text{ctx}}, z; \theta)$ from noisy mel frames x_t, conditioned on reference mel-spectrogram frames x_{ctx} and a frame-level phone sequence z. At inference, speech is generated by integrating

$$\frac{d\phi_t}{dt} = v_t(\phi_t, x_{\text{ctx}}, z; \theta), \qquad \phi_0 = x_0 \sim \mathcal{N}(0, I), \tag{1}$$

from t=0 to t=1. In ZS-TTS, x_{ctx} contains the reference prompt that specifies the speaker's voice.

At the Transformer input, the three frame-wise sequences (x_t, x_{ctx}, and the embedded phones) are concatenated and mapped by a single projection W_p; the prompt thus enters the model only through the block of W_p applied to the context frames $x^i_{\text{ctx}} \in \mathbb{R}^F$, which we denote W_{ctx}, where F is the mel dimension. Each context frame contributes

$$h^i = W_{\text{ctx}}\, x^i_{\text{ctx}} \in \mathbb{R}^D \tag{2}$$

to the Transformer input, where D is the embedding dimension. We call h^i the *conditioning features*: the prompt speaker's voice reaches the rest of the model only through them.

Deployment Setting and Scope ESIAN operates on a deployed ZS-TTS model with the eraser and gate enabled. Users may attempt to reproduce a target speaker's voice by submitting text and reference recordings of that speaker. We assume that users cannot modify the model or disable the eraser and gate. Our evaluation measures suppression using reference recordings not used for enrollment.

Problem Formulation We formalize the task of erasing speaker identity in the continual setting, where requests arrive over time rather than at once (Lee et al., 2026a; Xu et al., 2026). The k-th

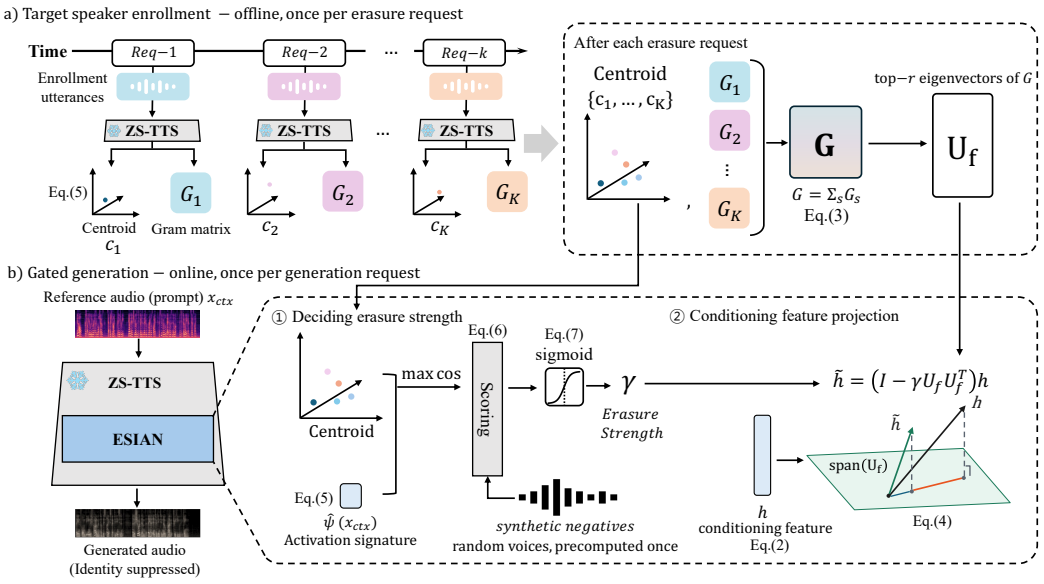

Figure 2: Overview of ESIAN. A removal request supplies one or more speakers, and each contributes one Gram matrix G_s and one centroid c_s to the accumulated state; the illustration draws a single speaker per request. The gate reads $\hat{\psi}$ from the Transformer blocks in one extra forward pass; the eraser acts on h at the input projection during generation.

request supplies a set of speakers $\mathcal{S}_k$, each represented by a set of utterances E_s, the speaker's *enrollment set*. We assume no access to the original training corpus, the training pipeline, or remaining speakers' utterances. We write $\mathcal{F}_k = \bigcup_{j \le k} \mathcal{S}_j$ for the target set accumulated after k requests, and $K = |\mathcal{F}_k|$ for the number of target speakers. We consider three objectives, which extend the erasure and preservation criteria of Kim et al. (2025) to sequential requests.

(i) **Erasure**: after each request k, suppress reproduction of voices from the accumulated target set $\mathcal{F}_k$, without relying on an external oracle that flags which prompts belong to a target speaker.

(ii) **Preservation**: preserve the model's original voice-cloning capability for remaining speakers outside the target set.

(iii) **Continual update**: incorporating a new request should not require revisiting utterances from earlier requests, which the provider may no longer retain.

4 Method

The prompt acts on the model only through the conditioning features (Section 3). ESIAN applies a closed-form projection computed from summed per-speaker statistics to suppress target-speaker identity during generation. This *eraser* is not speaker-selective, so a *gate* uses the model's internal signals to set its strength for each prompt. Enrollment collects eraser statistics and gate signatures from the frozen backbone without retaining the audio. Figure 2 summarizes enrollment and generation; the supplementary material provides the procedure.

4.1 Closed-Form Speaker Eraser

The eraser first identifies the directions to remove. Each enrollment utterance of speaker s is supplied as the context x_{ctx}, and the conditioning features h^i it produces at the input projection (Eq. 2) are stacked row-wise into a matrix $H_s \in \mathbb{R}^{N_s \times D}$, where N_s is the number of enrollment frames of that speaker. The Gram matrix $G_s = H_s^\top H_s \in \mathbb{R}^{D \times D}$ records how much each direction is used across

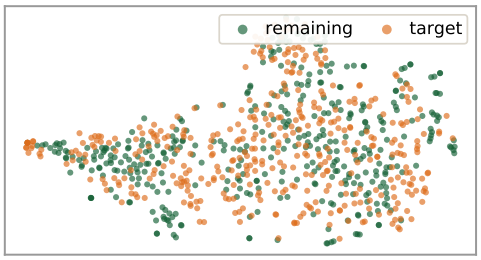

Figure 3: t-SNE of conditioning features h for target and remaining speakers. The two populations overlap substantially in this visualization.

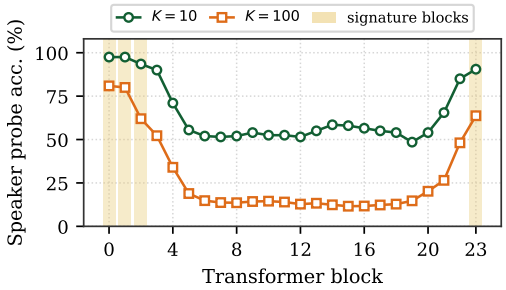

Figure 4: Speaker-classification accuracy of linear probes across Transformer blocks. Probes classify among the K enrolled target speakers; chance level is $1/K$. Shaded blocks form the activation signature.

those frames, and each request accumulates the matrices of its own speakers into a running sum,

$$G = \sum_{j \le k} \sum_{s \in \mathcal{S}_j} G_s = \sum_{s \in \mathcal{F}_k} G_s. \tag{3}$$

The accumulated sum is independent of request order or grouping, so the eraser after k requests is the same as if all k requests were processed at once (supplementary material). Continued suppression of earlier target speakers is evaluated empirically. The eigendecomposition $G = Q\Lambda Q^\top$, with eigenvalues in decreasing order, ranks directions by how much of the target speakers' conditioning features they account for, and the first r columns of Q, denoted $U_f \in \mathbb{R}^{D \times r}$, span the dominant ones. Removing them removes most of what a prompt from those speakers contributes (Section 5.3).

Erasure removes those components from every conditioning frame,

$$\tilde{h}^i = (I - \gamma U_f U_f^\top)\, h^i, \tag{4}$$

where $\gamma \in [0, 1]$ controls the erasure strength: $\gamma = 0$ leaves h^i unchanged, and $\gamma = 1$ removes all components spanned by U_f. The gate of Section 4.2 computes it for each incoming prompt. The edit costs one matrix multiplication per frame, and it leaves the weights, the phone inputs, and the noisy frames untouched.

The construction would be complete if U_f were specific to the target speakers. In practice U_f is not: measured on remaining speakers, 99.9% of the squared norm of the conditioning features already lies within the same r directions, so the projection strips a remaining speaker almost as thoroughly as a target speaker (Figure 3). This overlap in the selected subspace motivates adjusting the erasure strength for each prompt. The gate reads activations at other points in the network, where linear-probe analysis reveals speaker identity information (Figure 4), and uses them to distinguish target from remaining prompts (Figure 5).

4.2 Intrinsic Gating

The intrinsic gate maps each incoming context x_{ctx} to an erasure strength γ based on its activation similarity to enrolled speakers, without an external verifier.

Speaker enrollment Enrollment summarizes each target speaker as a single vector for comparison with incoming prompts. Linear-probe analysis shows that speaker classification is most accurate at the shallowest and deepest blocks (Figure 4), supporting the choice of $\mathcal{L} = \{0, 1, 2, 23\}$. Activations from these blocks are collected in a single forward pass with x_{ctx} alone. The gate pools activations over time and concatenates them across blocks into an utterance-level signature ψ. Whitening rescales the signatures using the enrollment covariance, and each speaker is represented

by a centroid of the whitened signatures for comparison with incoming prompts,

$$\begin{aligned} \psi &= \big[\,\mathrm{mean}_t\, a_t^{(\ell)}\;;\;\mathrm{std}_t\, a_t^{(\ell)}\,\big]_{\ell\in\mathcal{L}}, \\ \hat{\psi} &= \nu\big(W(\psi-\bar{\psi})\big), \\ c_s &= \nu\big(\textstyle\sum_{u\in E_s}\hat{\psi}(u)\big), \end{aligned} \tag{5}$$

where $a_t^{(\ell)}$ is the activation of block ℓ at frame t, $\nu(v) = v/\|v\|$, $\bar{\psi}$ is the enrollment mean, and W is the inverse square root of the enrollment covariance. The component ablation supports the use of whitening for speaker discrimination (Appendix D).

Scoring and normalization At inference time, a prompt is scored by the largest similarity to any enrolled speaker's centroid, $S(x_{\mathrm{ctx}}) = \max_s \cos\big(\hat{\psi}(x_{\mathrm{ctx}}), c_s\big)$. Interpreting S requires a reference distribution for non-target prompts. Since remaining-speaker recordings are unavailable, we construct this reference from N prompt-free generations of the model. Their activation signatures form a *synthetic-negative bank* for score normalization and threshold calibration. Following adaptive score normalization (Matejka et al., 2017), we normalize the raw score using the prompt's similarities to the top-scoring synthetic negatives,

$$g(x_{\mathrm{ctx}}) = \frac{S(x_{\mathrm{ctx}}) - \mu_C(x_{\mathrm{ctx}})}{\sigma_C(x_{\mathrm{ctx}})}, \tag{6}$$

where μ_C and σ_C are the mean and standard deviation of the similarities between x_{ctx} and its top-scoring synthetic negatives. The normalization expresses every score in standard deviations above the prompt's own cohort, so that one threshold serves every prompt.

Threshold calibration The normalized score is compared against a threshold τ: too low and remaining speakers are wrongly erased, too high and target speakers are missed. Rather than tuning τ for performance, we fix a budget for wrongful erasure and derive τ from it. We use synthetic-negative scores as a proxy for those of remaining-speaker prompts and evaluate the resulting calibration on held-out remaining speakers (Section 5.3). To account for repeated prompts from the same speaker, we specify a target false-positive budget α over M prompts per speaker. Allocating α/M to each prompt sets τ to the empirical $(1-\alpha/M)$ quantile of the synthetic-negative scores.

Erasure strength The score and the threshold set how much of the projection each prompt receives (Section 4.1), through a soft gate,

$$\gamma = \sigma(\kappa\,(g-\tau))\,. \tag{7}$$

Near the threshold, a binary gate risks leaving target prompts unerased or fully erasing remaining prompts. The sigmoid instead varies erasure strength smoothly around τ, with κ controlling the transition width. Threshold decisions remain unchanged across the tested positive κ values (supplementary material). Unlike refusal, which rejects falsely flagged remaining requests entirely, ESIAN returns speech for the requested text while adjusting erasure strength.

5 Experiments

5.1 Experimental Setup

Datasets Target speakers are drawn from LibriHeavy (Kang et al., 2024), a 50,000-hour English audiobook corpus, and remaining speakers from LibriSpeech test-clean (Panayotov et al., 2015), which does not overlap with it. The main comparison (Table 1) follows the original TGU (Kim et al., 2025) in its target speakers and evaluation setup. The continual experiment (Table 2) uses a balanced protocol: speakers arrive in subsequent requests up to $K = 100$, and every target speaker keeps the same enrollment budget and number of evaluation pairs. Enrollment and evaluation utterances are disjoint throughout. The main and balanced protocols use 3 s and 5 s prompts, respectively, and a budget of 40 enrollment utterances per speaker (Appendix A).

Table 1: Comparison under the TGU protocol. Train-free: no backbone updates; Ext.-free: no external detector. TruS† applies steering to all prompts in VoiceBox. Cost denotes training or approximate single-speaker enrollment compute; ESIAN excludes I/O and gate refitting (Appendix B). Ground Truth is the WER reference; dashes denote unreported TruS cost or inapplicable reference entries.

	SIM-F ↓		SIM-R ↑		WER ↓		Train-	Ext.-	
Method	micro	macro	micro	macro	F	R	free	free	Cost
Original	0.704	0.576	0.615	0.616	2.39	7.75	–	–	–
Ground Truth	–	–	–	–	1.43	2.29	–	–	–
TruS† (Lee et al., 2026b)	0.686	0.579	0.596	0.596	2.32	7.83	✓	✓†	–
TGU (Kim et al., 2025)	0.171	0.148	0.594	0.596	2.37	7.89	×	✓	∼17 h
SGU (Kim et al., 2025)	**0.158**	0.148	0.447	0.457	2.24	6.88	×	✓	∼5 h
ESIAN (Ours)	0.196	**0.141**	**0.604**	**0.609**	2.27	7.75	✓	✓	∼2.5 s

Baselines Original denotes unmodified VoiceBox (Le et al., 2023). TGU finetunes the model using a teacher that redirects target-speaker prompts toward random voices, while SGU finetunes without a teacher (Kim et al., 2025). TruS suppresses speaker identity by steering hidden activations (Lee et al., 2026b). Our VoiceBox adaptation applies shared steering to all prompts without an external detector (Table 1); a separate oracle run uses target-speaker labels and per-speaker steering. We also evaluate an external ECAPA-TDNN gate (Desplanques et al., 2020) coupled to the same ESIAN eraser (Table 3).

Implementation Unless stated otherwise, we use $r = 40$ eraser directions and $N = 400$ synthetic negatives, generated once and reused by every subsequent request, and the threshold is calibrated with $\alpha = 0.15$ and a prompt budget of $M = 20$. Whitening uses shrinkage 0.1, normalization uses the top 100 synthetic negatives, and the nominal slope is $\kappa = 35$. The main comparison shares the generation settings: 31 midpoint ODE steps, a guidance scale of 0.7, and a HiFi-GAN vocoder (Kong et al., 2020). Full settings and stored-run details appear in Appendix B.

Evaluation Metrics Speaker similarity (SIM) is the cosine similarity between WavLM-TDCNN (Chen et al., 2022) embeddings of generated speech and its prompt; lower target SIM indicates stronger erasure, while remaining SIM should stay close to Original. Unequal target-pair counts motivate reporting both pair-level (micro) and equally weighted speaker-level (macro) averages. Word error rate (WER) compares HuBERT-L (Hsu et al., 2021) transcriptions with the requested text. Original provides the pre-intervention baseline, while Ground Truth indicates transcription difficulty on the corpus recordings.

5.2 Evaluation

Table 1 evaluates target-speaker suppression, remaining-speaker preservation and content accuracy. Our erasure reaches 0.196 on the micro average, against 0.171 for TGU and 0.158 for SGU, all far below the Original's 0.704. Remaining similarity stays close to Original (0.604 versus 0.615). Paired speaker-level bootstrap intervals support the preservation gain over TGU; the difference from TruS is not statistically resolved (Appendix I). ESIAN updates no backbone weights and calls no external detector. Its approximate single-speaker enrollment compute is 2.5 s, excluding preprocessing and gate refitting, versus reported training costs of 17 h for TGU and 5 h for SGU. At generation time the gate adds one forward pass, 32 ms, under 1% of measured synthesis time (Appendix B).

TruS reaches SIM-F 0.686 under the main-table adaptation. With oracle labels, a separate run reaches 0.517 on the balanced target list, using per-speaker steering, strength 1.2, and a 1,500-frame target cap. These different conditions make it a separate reference rather than a matched oracle ablation (Appendix A).

The micro and macro averages differ because one speaker supplies 62% of the target pairs and dominates the micro number. Per speaker the three are comparable, 0.141 against 0.148 for both TGU and SGU. The gate drives 94.7% of target prompts to a mean similarity of 0.171, the level Kim

Table 2: Sequential requests under the balanced protocol. Each SIM-F cell tracks the speakers introduced in the column task after the row task. Entries are request-group averages; individual prompt tracking is reported in Appendix G.

			SIM-F of task-j speakers ↓					
Task	Added	Total	1	2	3	4	5	SIM-R ↑
1	10	10	0.179	–	–	–	–	0.647
2	+10	20	0.167	0.179	–	–	–	0.636
3	+10	30	0.167	0.181	0.174	–	–	0.642
4	+20	50	0.168	0.187	0.178	0.190	–	0.648
5	+50	100	0.179	0.203	0.178	0.202	0.187	0.648

et al. (2025) associate with pairs of different real speakers; the remaining 5.3% pass the gate and keep 0.641. We refer to the fraction below the gate threshold as *leak*. These are residual suppression failures; Appendix H reports their concentration across speakers.

Despite target SIM falling from 0.704 to 0.196, target WER remains 2.27 versus Original's 2.39, and remaining WER matches Original at 7.75, supporting identity suppression with preserved content accuracy. On the 33 falsely flagged remaining pairs, ESIAN also returns speech with WER 5.34%, versus 5.49% for Original, despite altered voice identity (Appendix D). Erasing therefore retains spoken-content delivery on these requests, whereas refusing the same flagged requests would provide no audio. To check whether identity suppression accompanies audio degradation, we evaluate UTMOS (Saeki et al., 2022), an automatic quality predictor. Remaining outputs score 4.03, matching Original, while target outputs score 4.13 versus Original's 3.63 (Appendix I).

Continual erasing Across five sequential requests enrolling 100 speakers, each update accumulates the Gram matrix, recomputes the eigenspace, refits whitening and centroids from stored signatures, and recalibrates the threshold on the synthetic bank. Table 2 tracks each request group through subsequent enrollments and reports remaining-speaker preservation. Average erasure remains stable as the target set grows tenfold: request-group SIM-F lies between 0.167 and 0.203, far below Original's 0.63, and remaining similarity stays at 0.636–0.648. The first request's speakers stay at 0.167–0.179. At $K = 100$, the 50 previously enrolled speakers have a median SIM-F change of $+0.002$ from enrollment (10th–90th percentiles $[-0.003, +0.037]$). Appendix G provides all individual trajectories and gate-decision changes.

To check whether outputs instead resemble *other* enrolled speakers, we also compute their maximum similarity to the entire target set (FSSIM). At $K = 100$, the mean is 0.412, close to 0.415 for remaining outputs scored against the same speaker references; Appendix H.1 reports the full sweep and reference distributions.

Out-of-domain speakers Speakers absent from the backbone's training data may also request that their voices no longer be cloned. We evaluate ten such speakers from LibriTTS (Zen et al., 2019), constructing the eraser and gate solely from their enrollment recordings. SIM-F falls from 0.672 to 0.169, comparable to the in-domain value of 0.179 under the balanced protocol (Table 2). SIM-R remains at 0.651 versus Original's 0.658 and target-side WER is 4.4%. Without domain-specific threshold adjustment, the observed false-positive rate is 1.4%, with no leakage in this evaluation (Appendix J).

5.3 Ablation Study

All ablations use the balanced protocol of the continual experiments.

Gate necessity Consistent with the conditioning-feature overlap, increasing uniform erasure strength reduces both target and remaining similarity; none of the tested strengths combines strong suppression with preservation (Table 3). At $\gamma = 1$, SIM-F/R is $0.163/0.154$; intrinsic gating gives $0.179/0.647$. The intrinsic gate thus retains strong target suppression while preserving remaining-speaker similarity close to Original's 0.658. An oracle gate applies the same eraser only to target

Table 3: Uniform versus gated projection under the balanced $K = 10$ protocol. All rows use the same backbone; intervention rows use the top-40 eraser. Only ECAPA uses an external speaker verifier for gating.

Projection strength / gate	SIM-F ↓	SIM-R ↑
$\gamma = 0$ (Original)	0.631	0.658
$\gamma = 0.2$	0.613	0.630
$\gamma = 0.6$	0.447	0.471
$\gamma = 1$	0.163	0.154
Oracle gate	0.163	0.658
External ECAPA gate	0.165	0.658
Intrinsic gate (ESIAN)	0.179	0.647

prompts using ground-truth speaker labels. The intrinsic gate approaches this reference, with differences of 0.016 in SIM-F and 0.011 in SIM-R.

External gate and projection controls We compare intrinsic and external gating with the same ESIAN eraser to evaluate speaker-selective erasing using backbone-derived signatures. At $K = 10$, intrinsic and ECAPA gating yield SIM-F/R of $0.179/0.647$ and $0.165/0.658$, respectively (Table 3). At $K = 100$, the corresponding scores are $0.190/0.648$ and $0.196/0.648$ (Appendix F, Table 10). These results demonstrate selective erasing with the model's internal signatures without an additional speaker verifier. With the intrinsic gate fixed, 40 random projection directions give SIM-F 0.387, versus 0.179 for top-40 and 0.180 for the full conditioning span. The selected rank captures 99.3–99.4% of enrollment energy across $K = 10$–100 (Appendix E).

Calibration validity This evaluation checks whether calibration on synthetic negatives transfers to real non-target speech. On two held-out sets of 40 speakers each, at least one prompt exceeds the calibrated threshold for 4 remaining speakers (10%) and 3 non-target LibriHeavy speakers (7.5%). The LibriHeavy set additionally tests non-target speakers from the same corpus as the target speakers. Both observed rates are below the 15% speaker-level budget, supporting synthetic-negative calibration on these sets (Appendix D.1).

Gate design Enrollment amount affects detection coverage. As enrollment grows from 5 to 40 utterances per speaker, leak falls from 74% to 3.5% at $K = 10$ and from 15% to 4.5% at $K = 100$ (Figure 6); 40 is the enrollment size used throughout. The remaining sensitivity studies and aggregation comparisons appear in Appendices D and E.

6 Conclusion

We presented ESIAN, a training-free framework for speaker identity erasure in ZS-TTS. A closed-form eraser acts on conditioning features, and an intrinsic gate sets its strength for each prompt. Requests are incorporated through accumulated statistics without revisiting earlier audio. The evaluated outputs show target-speaker suppression with preserved text accuracy and remaining-speaker quality. Average suppression remains stable as the target set grows to a hundred speakers; individual prompt behavior is tracked separately.

7 Limitations

The evaluated operating point uses 40 enrollment utterances per speaker; smaller enrollment sets increase leakage (Figure 6). Evaluation uses VoiceBox; transfer to other conditioning architectures remains outside the validated setting. Protection assumes the deployed intervention remains enabled, and the held-out evaluation does not cover adaptive prompt selection or adversarial recording transformations. Individual speaker trajectories complement average continual performance (Appendix G); full update-cost scaling, including gate refitting, remains to be measured.

AI USE STATEMENT

In this work, generative AI tools were used to assist with drafting and revising parts of the manuscript, including improvements to readability, phrasing, and organization. They also supported literature discovery, research discussions, code implementation, and experimental analysis. No LLM-generated synthetic datasets were used; the synthetic speech for gate calibration was generated by the TTS backbone as described in the method. The authors are responsible for reviewing and verifying AI-assisted text, code, and analyses, and take full responsibility for the scientific claims and final manuscript.

ETHICS STATEMENT

This work is motivated by concerns about unauthorized voice cloning and its potential misuse for impersonation. ESIAN provides a mechanism for suppressing enrolled speakers' identities during speech generation while preserving spoken content and voice cloning for remaining speakers. Its intended use is to support speaker requests within a provider-controlled inference pipeline.

Our experiments use existing recordings from publicly available speech corpora, including Libri-Heavy, LibriSpeech, and LibriTTS. No new recordings were collected, and no human-participant studies were conducted. Speakers are identified by dataset speaker IDs. Subsequent enrollment updates use stored statistics without requiring earlier enrollment recordings to be revisited.

REPRODUCIBILITY STATEMENT

Section 4 describes the proposed method, and Appendix C provides pseudocode and the assumptions and proof for sequential–batch equivalence of eraser construction. Section 5.1 and Appendix A specify the datasets, evaluation protocols, baseline configurations, and enrollment–evaluation separation. Appendix B documents hyperparameters, random seeds, software versions, hardware, and computational cost measurements. Appendices D–J provide additional experimental settings, ablations, and detailed results, including external-gate comparisons and individual-speaker tracking. Appendix I describes the bootstrap procedures used to estimate uncertainty.

REFERENCES

Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: perfect linear concept erasure in closed form. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/d066d21c619d0a78c5b557fa3291a8f4-Abstract-Conference.html.

Lucas Bourtoule, Varun Chandrasekaran, Christopher A. Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *42nd IEEE Symposium on Security and Privacy, SP 2021, San Francisco, CA, USA, 24-27 May 2021*, pp. 141–159. IEEE, 2021. doi: 10.1109/SP40001.2021.00019. URL https://doi.org/10.1109/SP40001.2021.00019.

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 J. Sel. Top. Signal Process.*, 16(6):1505–1518, 2022. doi: 10.1109/JSTSP.2022.3188113. URL https://doi.org/10.1109/JSTSP.2022.3188113.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching. *arXiv preprint arXiv:2410.06885*, 2024.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. ECAPA-TDNN: Emphasized channel attention, propagation and aggregation in TDNN based speaker verification. In *Interspeech*, pp. 3830–3834, 2020. doi: 10.21437/Interspeech.2020-2650.

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pp. 2426–2436. IEEE, 2023. doi: 10.1109/ICCV51070.2023.00230. URL https://doi.org/10.1109/ICCV51070.2023.00230.

Chongyang Gao, Lixu Wang, Kaize Ding, Chenkai Weng, Xiao Wang, and Qi Zhu. On large language model continual unlearning. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/fc28053a08f59fccb48b11f2e31e81c7-Paper-Conference.pdf.

Aditya Golatkar, Alessandro Achille, and Stefano Soatto. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pp. 9301–9309. Computer Vision Foundation / IEEE, 2020. doi: 10.1109/CVPR42600.2020.00932. URL https://openaccess.thecvf.com/content_CVPR_2020/html/Golatkar_Eternal_Sunshine_of_the_Spotless_Net_Selective_Forgetting_in_Deep_CVPR_2020_paper.html.

Feng Han, Kai Chen, Chao Gong, Zhipeng Wei, Jingjing Chen, and Yu-Gang Jiang. Dumo: Dual encoder modulation network for precise concept erasure. In Toby Walsh, Julie Shah, and Zico Kolter (eds.), *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pp. 3320–3328. AAAI Press, 2025. doi: 10.1609/AAAI.V39I3.32343. URL https://doi.org/10.1609/aaai.v39i3.32343.

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 Trans. Audio Speech Lang. Process.*, 29:3451–3460, 2021. doi: 10.1109/TASLP.2021.3122291. URL https://doi.org/10.1109/TASLP.2021.3122291.

Wei Kang, Xiaoyu Yang, Zengwei Yao, Fangjun Kuang, Yifan Yang, Liyong Guo, Long Lin, and Daniel Povey. Libriheavy: A 50,000 hours ASR corpus with punctuation casing and context. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pp. 10991–10995. IEEE, 2024. doi: 10.1109/ICASSP48485.2024.10447120. URL https://doi.org/10.1109/ICASSP48485.2024.10447120.

Taesoo Kim, Jinju Kim, Dongchan Kim, Jong Hwan Ko, and Gyeong-Moon Park. Do not mimic my voice : Speaker identity unlearning for zero-shot text-to-speech. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*, pp. 30176–30198. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/kim25l.html.

Sangamesh Kodge, Gobinda Saha, and Kaushik Roy. Deep unlearning: Fast and efficient gradient-free class forgetting. *Trans. Mach. Learn. Res.*, 2024, 2024. URL https://openreview.net/forum?id=BmI5p6wBi0.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/c5d736809766d46260d816d8dbc9eb44-Abstract.html.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, and Wei-Ning Hsu. Voicebox: Text-guided multilingual universal speech generation at scale. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/2d8911db9ecedf866015091b28946e15-Abstract-Conference.html.

Justin Lee, Zheda Mai, Jinsu Yoo, Chongyu Fan, Cheng Zhang, and Wei-Lun Chao. Continual unlearning for text-to-image diffusion models: A regularization perspective. In *The Fourteenth International Conference on Learning Representations, ICLR 2026*. OpenReview.net, 2026a. URL https://openreview.net/forum?id=BsY20r9FOM. arXiv:2511.07970.

Myungjin Lee, Eunji Shin, and Jiyoung Lee. Erasing your voice before it's heard: Training-free speaker unlearning for zero-shot text-to-speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2026, Barcelona, Spain, May 4-8, 2026*, pp. 17627–17631. IEEE, 2026b. doi: 10.1109/ICASSP55912.2026.11462017.

Tae-Young Lee, Sundong Park, Minwoo Jeon, Hyoseok Hwang, and Gyeong-Moon Park. ESC: erasing space concept for knowledge deletion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 5010–5019. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.00472. URL https://openaccess.thecvf.com/content/CVPR2025/html/Lee_ESC_Erasing_Space_Concept_for_Knowledge_Deletion_CVPR_2025_paper.html.

Thanathai Lertpetchpun, Thanapat Trachu, Sai Praneeth Karimireddy, and Shrikanth Narayanan. Speech generation speaker poisoning: Capability erasure in zero-shot text-to-speech. *arXiv preprint arXiv:2603.07551*, 2026.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/forum?id=PqvMRDCJT9t.

Alessandro Mantelero. The EU proposal for a general data protection regulation and the roots of the "right to be forgotten". *Computer Law & Security Review*, 29(3):229–235, 2013. doi: 10.1016/j.clsr.2013.03.010.

Pavel Matejka, Ondrej Novotný, Oldrich Plchot, Lukás Burget, Mireia Díez Sánchez, and Jan Cernocký. Analysis of score normalization in multilingual speaker recognition. In Francisco Lacerda (ed.), *18th Annual Conference of the International Speech Communication Association, Interspeech 2017, Stockholm, Sweden, August 20-24, 2017*, pp. 1567–1571. ISCA, 2017. doi: 10.21437/INTERSPEECH.2017-803. URL https://doi.org/10.21437/Interspeech.2017-803.

Thanh Tam Nguyen, Thanh Trung Huynh, Zhao Ren, Phi Le Nguyen, Alan Wee-Chung Liew, Hongzhi Yin, and Quoc Viet Hung Nguyen. A survey of machine unlearning. *ACM Trans. Intell. Syst. Technol.*, 16(5):108:1–108:46, 2025. doi: 10.1145/3749987. URL https://doi.org/10.1145/3749987.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2015, South Brisbane, Queensland, Australia, April 19-24, 2015*, pp. 5206–5210. IEEE, 2015. doi: 10.1109/ICASSP.2015.7178964. URL https://doi.org/10.1109/ICASSP.2015.7178964.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pp. 7237–7256. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.ACL-MAIN.647. URL https://doi.org/10.18653/v1/2020.acl-main.647.

Shauli Ravfogel, Michael Twiton, Yoav Goldberg, and Ryan Cotterell. Linear adversarial concept erasure. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 18400–18421. PMLR, 2022. URL https://proceedings.mlr.press/v162/ravfogel22a.html.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. UTMOS: UTokyo-SaruLab system for VoiceMOS challenge 2022. In *Interspeech*, pp. 4521–4525, 2022. doi: 10.21437/Interspeech.2022-439.

Weijia Shi, Jaechan Lee, Yangsibo Huang, Sadhika Malladi, Jieyu Zhao, Ari Holtzman, Daogao Liu, Luke Zettlemoyer, Noah A. Smith, and Chiyuan Zhang. MUSE: machine unlearning six-way evaluation for language models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=TArmA033BU.

Kartik Thakral, Tamar Glaser, Tal Hassner, Mayank Vatsa, and Richa Singh. Continual unlearning for foundational text-to-image models without generalization erosion. *arXiv preprint arXiv:2503.13769*, 2025.

Anvith Thudi, Gabriel Deza, Varun Chandrasekaran, and Nicolas Papernot. Unrolling SGD: understanding factors influencing machine unlearning. In *7th IEEE European Symposium on Security and Privacy, EuroS&P 2022, Genoa, Italy, June 6-10, 2022*, pp. 303–319. IEEE, 2022. doi: 10.1109/EUROSP53844.2022.00027. URL https://doi.org/10.1109/EuroSP53844.2022.00027.

Paul Voigt and Axel von dem Bussche. *The EU General Data Protection Regulation (GDPR): A Practical Guide*. Springer, 2017.

Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*, 2023.

Xiaoyu Xu, Minxin Du, Kun Fang, Yaxin Xiao, Zhicong Huang, Cheng Hong, Qingqing Ye, and Haibo Hu. FIT to forget: Robust continual unlearning for large language models. *arXiv preprint arXiv:2601.21682*, 2026.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. LibriTTS: A corpus derived from LibriSpeech for text-to-speech. In *Interspeech*, pp. 1526–1530, 2019. doi: 10.21437/Interspeech.2019-2441.

A Dataset and Protocol Details

Datasets and speakers. The ten target speakers are drawn from LibriHeavy (Kang et al., 2024); the remaining speakers are the 40 speakers of the LibriSpeech test-clean set (Panayotov et al., 2015), and the two sets are disjoint.

Protocols. The main comparison (Table 1) reproduces the TGU protocol: the ten target speakers of Kim et al. (2025) are enrolled, prompts are cropped to their last 3 s (300 frames), and the evaluation pair list supplies 1,193 target and 1,235 remaining generation pairs; the target pairs are imbalanced, one speaker contributing 62%, so we report both micro and macro averages. The continual experiment (Table 2) uses a balanced protocol, so named because it removes that imbalance: every target speaker supplies exactly 20 evaluation pairs, so a request set of K speakers contributes $20K$ target pairs (200, 600 and 2,000 at $K = 10$, 30 and 100) against 800 remaining pairs from the 40 remaining speakers. Equal pairs per speaker make the micro and macro averages identical, which is why Table 2 reports a single number. Every speaker also keeps the same enrollment budget, prompts are cropped to their first 5 s, and speakers arrive over five sequential requests up to $K = 100$.

Averaging unit. Prior work in this setting does not state whether speaker similarity is averaged over evaluation pairs or over speakers. Kim et al. (2025) state the averaging rule only for spk-ZRF, a mean over samples, and their released evaluation code averages the per-pair similarities with no speaker grouping; Lee et al. (2026b) defer to that protocol, and Lertpetchpun et al. (2026) name the generated sample as the unit without saying how samples are combined. None of the three reports a per-speaker breakdown or any variation across speakers. We therefore read the pair-level (micro) average as the number comparable to published results, and the speaker-level (macro) average as a check that no single speaker carries the result, which the imbalance of the target pair list makes necessary (Appendix H).

Baseline configuration. TGU and SGU use the released unlearned checkpoints of Kim et al. (2025); we do not retrain them. TruS was developed for F5-TTS (Chen et al., 2024) and is adapted here to VoiceBox. The main-table run applies shared steering to all prompts with a threshold computed over the target set. The oracle run instead uses target-speaker labels and per-speaker layer/step selection, with steering strength 1.2 (0.5173 SIM-F; the strength-1.0 run gives 0.528). It uses the balanced 200-pair target list, but its historical implementation caps the target at 1,500 frames. This duration difference has not been isolated by a matched rerun. Consequently, the 0.686 and 0.517 results are neither the same-protocol oracle ablation nor a measurement of the oracle's isolated contribution. Re-evaluating all 1,193 saved waveforms of the corrected main-table TruS run with the same HuBERT pipeline gives WER-F 2.32%.

Enrollment. Each enrolled speaker supplies 40 utterances, 400 in all for the ten speakers of the main protocol. Eraser statistics and gate signatures are collected at the deployment prompt length, matching the condition scored at inference: three 3 s windows per utterance under the main protocol and 5 s under the balanced protocol. Appendix D shows that this coherence is necessary. Enrollment and evaluation utterances are disjoint throughout. Because the evaluation pairs are fixed by the protocol, an enrollment utterance that also appears in one of them is dropped from enrollment, which removes 12 of the 400 under the main protocol and leaves 388. The balanced protocol draws the two sets separately, and no overlap arises at any $K \in \{10, 20, 30, 50, 100\}$.

B Implementation Details

Constants. Table 4 lists every constant of the eraser, the gate and the generator. Two entries are consequences of the method rather than free choices: signatures are collected at the length the gate will score, for the reason Appendix D measures, and the negative bank is drawn from the model itself, so calibrating τ needs no remaining speech.

Runs, seeds and software. Each reported SIM value uses one generation per evaluation pair. The gate is deterministic for a fixed prompt and stored state, while synthesized audio depends on the generation seed. The reported uncertainty therefore reflects resampling of observed pairs or speakers, rather than variation over independent generation runs. Generation uses seed 1234 and bootstrap resampling seed 0. The out-of-domain speaker draw uses seed 0. Runs use Debian 12 with Python 3.9.25, PyTorch 2.1.0 (CUDA 11.8, cuDNN 8.7), torchaudio 2.1.0, transformers 4.35.0, SpeechBrain 1.1.0, jiwer 4.0.0, NumPy 1.26.4 and SciPy 1.13.1.

Cost. The approximate 2.5 s per-speaker compute estimate combines separately measured feature extraction, Gram accumulation, and eigendecomposition timings. It excludes audio loading, mel preprocessing, and gate refitting, including whitening, centroid updates, and threshold recalibration. This estimate does not establish full update-cost scaling through $K = 100$. The one-time synthetic bank is generated separately and reused. TGU and SGU costs in the main table are the previously reported training costs on two A100s, whereas ESIAN's estimate uses one A100; they are not matched-hardware speedup measurements. At inference, the gate adds 32.3 ms on average (median 32.7, standard deviation 1.1, 19 prompts), compared with 20.4 s for synthesis in the same run. This measured ratio is 0.16% and depends on target duration.

Table 4: Every constant of the eraser, the gate and the generator. Nominal settings are listed below; run-specific soft-gate slopes are documented in Appendix D; Appendix D measures how little the gate depends on α, M, N, κ, the block subset and the pooling.

Component	Setting
Eraser rank r	40
Gate blocks	$\{0, 1, 2, 23\}$ of 24
Pooling	mean and standard deviation over time
Signature length	three 3 s windows per utterance (main); first 5 s (balanced)
Whitening	from enrollment signatures, shrinkage 0.1
Score normalization	Adaptive score normalization, top 100 of $N = 400$ synthetic negatives
Threshold τ	$(1 - \alpha/M)$ quantile, $\alpha = 0.15$, $M = 20$
Transition width κ	35 (nominal; stored-run slopes below)
ODE solver	31 midpoint steps
Guidance scale	0.7
Precision	fp16
Vocoder	HiFi-GAN
Hardware	one A100 (40 GB); finetuned baselines unlearned on two

Algorithm 1 Enrollment and gated generation

once: sample N unconditional generations; store their signatures
procedure ENROLL($\mathcal{S}_k$, $\{E_s\}$)
for each speaker $s \in \mathcal{S}_k$ **do**
  collect conditioning features H_s from E_s; $G \leftarrow G + H_s^\top H_s$
  store the activation signatures of E_s
end for
$U_f \leftarrow$ top-r eigenvectors of G
refit the whitening on the stored signatures; set each centroid c_s by Eq. (5)
$\tau \leftarrow$ empirical $(1-\alpha/M)$ quantile of the stored negatives, rescored against the current centroids and whitening
procedure GENERATE(x_{ctx}, z)
$\psi \leftarrow$ pooled activations from one forward pass ($x_t = 0$ at $t = 0.5$)
$g \leftarrow$ adaptively normalized $\max_s \cos(\hat{\psi}, c_s)$, Eq. (6)
$\gamma \leftarrow \sigma(\kappa\,(g - \tau))$
return speech from Eq. (1) with $\tilde{h}^i = (I - \gamma U_f U_f^\top)\, h^i$

C Procedure and Continual Composition

C.1 Enrollment and generation

Algorithm 1 states the two procedures the main text describes: what a request does, and what a prompt does.

C.2 Sequential and batch construction

Persistent state. After request k, the system stores the accumulated Gram matrix, the enrolled speakers' activation signatures, the synthetic-negative bank, and the resulting whitening transform, centroids and threshold. The audio is not retained. Gate statistics are recomputed from the stored signatures; their storage scales with the number of retained signatures. With a fixed enrollment budget per speaker, this grows with the enrolled speaker count.

Proposition 1 *Fix the backbone and the enrollment sets $\{E_s\}_{s\in\mathcal{F}}$, and suppose the accumulated Gram matrix G has $\lambda_r > \lambda_{r+1}$. For any partition of $\mathcal{F}$ into requests $\mathcal{S}_1, \ldots, \mathcal{S}_m$ and any arrival order, the eraser and the gate obtained by processing the requests sequentially are identical to those obtained by processing $\mathcal{F}$ in a single request.*

Proof. For the eraser, $G_s = H_s^\top H_s$ depends only on speaker s's enrollment, and matrix addition is commutative and associative, so the accumulated G after the last request equals $\sum_{s\in\mathcal{F}} G_s$ regardless of the partition or order. U_f is a deterministic function of G: the assumed eigengap makes the projection $U_f U_f^\top$ unique, so the eraser is well defined and order-independent. For the gate, each speaker's signature set depends only on that speaker's enrollment; the whitening moments are sums over the union of the signature sets; the centroids are per-speaker means; and τ is the $(1-\alpha/M)$ quantile of the fixed negative bank scored against the current centroids and whitening. Each quantity is a deterministic function of the order-independent state, and therefore identical to the batch computation. □

The eigengap. The one assumption of Proposition 1 is the only place where continual composition could fail: a tie at $\lambda_r = \lambda_{r+1}$ leaves the top-r subspace undetermined, and two orders could then select different erasers from the same G. The assumption holds in every configuration we run: at $r = 40$ the relative gap $\lambda_{40}/\lambda_{41} - 1$ ranges from 5.0% to 9.7% across $K \in \{10, 20, 30, 50, 100\}$, and is 0.9% under the main protocol. Even that smallest gap is more than ten orders of magnitude above the precision at which the spectrum is resolved, so no configuration approaches the tie that would leave the top-r subspace undetermined. Recomputing the gate from sequentially accumulated and batch statistics with the synthetic-negative calibration gives centroid cosines above 0.9999999 and a maximum threshold difference of 1.25×10^{-8} across $K \in \{10, 20, 30, 50, 100\}$. These checks concern numerical equivalence of the constructed state; the behavior of earlier prompts is tracked separately in Appendix G.

Consequence. For fixed enrollment data and the stated eigengap condition, sequential and batch construction produce the same eraser and gate. This equality does not imply that individual outputs remain unchanged when additional speakers enter the target set.

D Gate Calibration and Enrollment

Operating points. The ablations in this section are measured on the main-protocol evaluation prompts subsampled with a per-speaker cap: at most 40 target prompts per speaker (177 over the nine speakers that contribute pairs) and 5 remaining prompts per speaker (200 over 40). Three quantities are reported. *Leak* is the fraction of target prompts scoring below τ, prompts that should have been erased and were not; *FP* is the fraction of remaining prompts above τ, prompts that should have been preserved and were not. *AUC* is the area under the ROC curve obtained by ranking target against remaining prompts by g. Leak and FP describe one operating point, the one τ selects, whereas AUC is independent of τ and describes how separable the two populations are at all. The two are reported separately because what follows turns on the distinction: α, M, N and κ move the operating point without changing the separation.

The leak here (4.0%) is lower than the end-to-end leak of the main text (5.3% over $1{,}193$ pairs) because of speaker imbalance rather than any difference in measurement. In the full pair list speaker 3912 supplies 62% of the pairs and has the highest leak rate of the nine (6.2%); the per-speaker cap reduces its share to 22.6%. The gap is the same phenomenon that separates the micro and macro averages of Appendix A.

Matching enrollment and deployment duration. The signature is a mean- and standard-deviation pooling of activations over time, and is therefore specific to the duration it is pooled over. Holding the whitening at the deployment length (3 s) and lengthening only the centroid enrollment windows collapses the gate: AUC falls from 0.997 (3 s) to 0.602 (5 s) and 0.563 (8 s), because a 3 s query signature no longer matches the same speaker's longer-window centroid; the raw match to its own centroid drops from 0.075 to 0.016, below a remaining prompt's 0.021. Re-fitting the whitening at the longer duration does not help (0.551 and 0.343): the whitening then no longer matches the 3 s queries and the score normalization degenerates. All gate statistics are therefore collected at the

Table 5: Gate components, one substitution at a time, on 177 target and 200 remaining prompts. Leak is reported at a matched 2.5% false-positive rate so that the rows are comparable.

Variant	AUC	Leak (%)
Deployed	**0.9969**	**4.0**
Whitening → centering and unit norm	0.8963	25.4
Adaptive score normalization → raw maximum cosine	0.9968	4.5
Pooling: mean only	0.9925	5.6
Pooling: standard deviation only	0.9936	4.0
Blocks $\{0\}$	0.9963	4.0
Blocks $\{23\}$	0.9971	2.3
Blocks $\{0, 1, 2\}$	0.9970	4.0
Blocks $\{0, 23\}$	0.9970	4.0
$N = 100$	0.9963	4.0
$N = 200$	0.9960	4.5
$N = 300$	0.9972	4.0

length the gate will score, 3 s windows under the main protocol and 5 s under the balanced protocol; at inference the incoming prompt is cropped to the same length before scoring, so the coherence is maintained by construction.

Three windows per utterance suffice. The question is how many signatures to take from one enrollment utterance. Each utterance contributes 3 s windows taken at its start, middle and end, with one signature computed per window, and a speaker's signatures are averaged into their centroid. One window per utterance (388 signatures) gives AUC 0.943 and 45.8% leak; three (1,158 signatures, a few utterances being too short for all three) give 0.997 and 4.0%; densifying to overlapping windows every 1.5 s or 1 s, up to 4,711 signatures, gives 0.996–0.997. The improvement supports sampling multiple temporal regions; denser overlapping windows yield little further change on this evaluation.

Gate components. Table 5 replaces one component of the gate at a time, holding everything else at the deployed setting. Removing whitening has the largest effect among the tested substitutions: AUC drops from 0.9969 to 0.8963, and leak rises from 4.0% to 25.4% at a matched 2.5% false-positive rate. This ablation supports whitening for speaker discrimination in the activation-signature space. The other substitutions change leak by at most about 1.7 percentage points on these 177 prompts.

Adaptive score normalization has little effect on separation in this ablation (AUC 0.9969 with normalization versus 0.9968 without it). It expresses each match in standard deviations relative to the prompt's selected synthetic negatives before threshold calibration. Pooling both moments retains information about activation level and temporal variation. The four blocks are those standing above the gap in the probe curve: at $K = 100$, where chance is 1%, blocks 0, 1, 23 and 2 classify the enrolled speakers at 80.9, 80.0, 63.7 and 62.0% while the next block reaches 52.2%. The rule is applied at the larger scale because at $K = 10$ chance is 10% and the top of the curve is flat, with four blocks within seven points of each other. No subset of those four is better than the full set by more than the resolution above, so the rule is a way of fixing the subset without tuning it rather than a claim that these four are optimal.

The size of the negative bank does not change separation either (0.9960–0.9972 from $N = 100$ to 400). We use 400 for the stability of the threshold rather than for separation. Since τ is the $(1 - \alpha/M) = 99.25\%$ quantile of the negative scores, N determines how many samples estimate that quantile: at $N = 100$ the top single score effectively fixes τ, at $N = 400$ the top three do. A threshold that moves when the bank is redrawn would undermine the budget rule itself, so where separation is flat we take the bank that estimates the quantile most stably. The bank is generated once and reused, so its cost is amortized.

The error budget. Table 6 sweeps α and M. The AUC is 0.9969 in every one of the six settings: the budget parameters choose how conservative the gate is, not how well it separates, and the two realized rates slide monotonically along one curve.

Table 6: The budget parameters. AUC is 0.9969 throughout; only the operating point moves. Bold marks the setting used everywhere.

α	M	FP (%)	Leak (%)
0.05	20	1.0	5.1
0.10	20	1.5	5.1
0.15	**20**	**2.5**	**4.0**
0.25	20	5.5	1.1
0.15	10	6.0	1.1
0.15	40	1.5	5.1

The budget parameters set the tradeoff: raising α from 0.15 to 0.25 takes leak from 4.0% to 1.1% and false positives from 2.5% to 5.5%. We report $\alpha = 0.15$ throughout as a target preservation budget; the achieved error rates are evaluated empirically.

Transition width and stored runs. We use $\kappa = 35$ as the nominal setting. The stored main-protocol generations used a score-derived slope of approximately 30.42, and historical balanced runs used slopes between approximately 14.7 and 35.19. Re-evaluating their stored scores with $\kappa = 35$ leaves threshold decisions unchanged, although soft strengths change. Likewise, sweeping positive slopes from 10 to 100 leaves the threshold boundary unchanged. This score analysis does not establish identical generated audio across slopes. The generation results in the paper refer to the saved runs, rather than to new generations at every tested slope.

Soft gate against a binary decision. The sigmoid assigns intermediate strengths near the threshold, unlike a binary gate. In the saved main-protocol outputs, 15 target and 13 remaining pairs satisfy $0.1 < \gamma < 0.9$. These counts describe the historical slope used in those runs. They identify prompts most affected by changing the transition width, but do not substitute for a generation comparison between hard and soft gating.

Erasure against refusal. A detector also supports a refusal policy. Among 1,235 remaining pairs in the main protocol, 33 score above the threshold; a policy rejecting all such prompts would return no audio for these requests. ESIAN instead returns speech, with mean SIM-R 0.248 on these pairs versus 0.628 for Original. The eight flagged pairs with partial erasure have mean SIM 0.379, while the 25 with $\gamma > 0.9$ have mean SIM 0.206. Re-evaluating all 33 flagged pairs with the main-table HuBERT pipeline, matching ESIAN and Original waveforms by source, target and transcript, gives mean per-pair WER 5.34% and 5.49%, respectively. All 66 waveforms are included. Thus the observed identity change is accompanied by retained transcription accuracy: these requests still receive the spoken text instead of no output. Both policies are compared on the same 33 requests flagged by the intrinsic gate. A soft gate also applies small nonzero strengths below threshold, so the two policies are not mathematically identical on all unflagged prompts.

Four prompts end to end. Table 7 follows one prompt from each of the four outcomes the gate produces. The target prompt the gate catches is erased to 0.169 from the 0.842 it would otherwise reach. The one it misses scores below τ, receives $\gamma \approx 0$, and is cloned as the original model would clone it, which is what makes the leak a detection miss rather than a failure of the eraser. On the remaining side the prompt the gate passes is generated unchanged to three decimals, and the false positive is erased almost in full.

D.1 Synthetic-negative checks and enrollment amount

The synthetic-negative bank supplies a proxy distribution for calibration. Figure 5 shows the scores at $K = 100$; Figure 6 varies enrollment amount under the synthetic-negative gate. At each amount the gate statistics are refitted, rather than retaining the whitening fit from 40 utterances. Five random enrollment subsets are evaluated per amount. Mean leak at 5, 10, 20 and 40 utterances is 73.7, 40.4, 18.1 and 3.5% for $K = 10$, and 15.3, 12.4, 6.5 and 4.5% for $K = 100$. No extrapolation beyond 40 utterances is made.

Table 7: One prompt from each gate outcome under the main protocol ($\tau = -0.543$). Each row is the pair whose similarity is closest to the median of its outcome group, so the rows are typical rather than selected; the groups hold 1,130, 63, 1,202 and 33 pairs. *Unerased* is the same pair generated by the original model.

Prompt	Speaker	g	γ	SIM	Unerased
Target, erased	3912	0.73	1.00	0.169	0.842
Target, leak	3912	−1.54	0.00	0.672	0.671
Remaining, preserved	1320	−2.39	0.00	0.621	0.621
Remaining, false positive	1580	−0.37	0.99	0.269	0.616

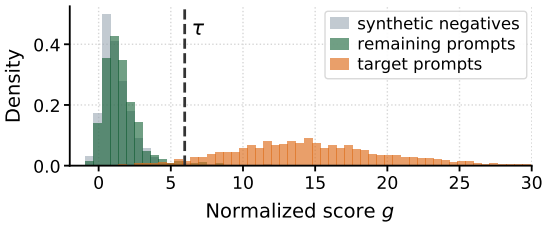

Figure 5: Normalized score g of synthetic negatives, remaining prompts, and target prompts against the calibrated threshold τ at $K = 100$.

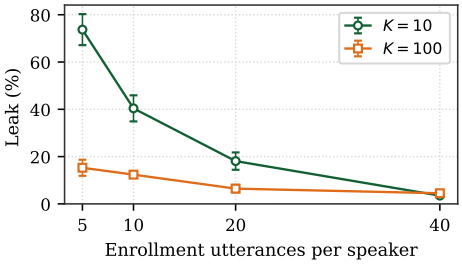

Figure 6: Leak, the fraction of target prompts that pass the gate, against the number of enrollment utterances per speaker, at $K = 10$ and $K = 100$. Points are means over five random draws of each speaker's enrollment subset, with bars at one standard deviation; at 40 utterances every utterance is used.

The speaker-level check flags $4/40$ held-out LibriSpeech speakers and $3/40$ disjoint LibriHeavy speakers at least once, versus a target budget of 15%. These are observed rates on the evaluated recordings, not a guarantee under arbitrary prompt selection or recording transformations. Both calibration and deployment performance depend on how well the generated negatives represent the score distribution of remaining speakers.

E Eraser and Design Controls

Centroid aggregation. A target prompt's evidence is concentrated in a single match: its similarity to its own centroid dominates, about four times a remaining prompt's closest match, while its similarity to every other enrolled speaker falls at or below zero (Figure 7a). The maximum over centroids separates the two populations in Figure 7b. We use this same aggregation for all prompts; at $K = 100$, it gives the lowest leak among the aggregations in the evaluated grid (Figure 9). Figure 8a–b visualizes this geometry on the enrollment signatures.

Eraser rank. The eraser keeps $r = 40$ eigendirections. Measured on the conditioning frames of the evaluation prompts, that subspace holds 99.3% of the target speakers' energy and 99.2% of the remaining speakers'; both reach 99.8% at $r = 60$ and saturate by $r = 80$, the rank of the conditioning pathway. Target and remaining prompts therefore occupy U_f to the same degree, which is the entanglement that makes the gate necessary. The figure quoted in the main text, 99.9% for remaining speakers, is this same quantity measured after each utterance is pooled into a single vector. Table 8 sweeps the rank on both protocols, with the gate held fixed so that only the projection changes. Too small a rank costs erasure on both protocols: at $r = 20$ the target similarity rises to 0.231 from 0.196 under the main protocol and to 0.209 from 0.174 under the balanced one. Beyond 40 neither protocol is harmed, and what a larger rank buys follows how much of the enrollment the top-r subspace already covers. The main protocol uses fewer enrollment frames than the balanced

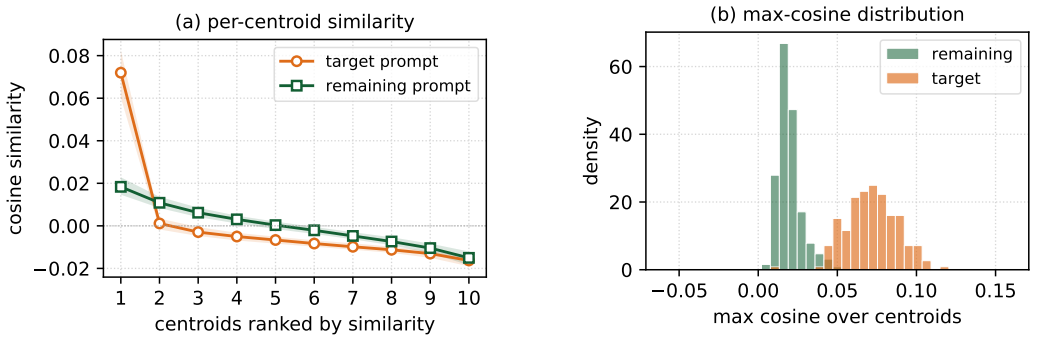

Figure 7: Centroid aggregation ($K = 10$). (a) A target prompt matches a single centroid strongly and all others near zero, while a remaining prompt scores low against every centroid. (b) The maximum over centroids separates the target and remaining score distributions in this evaluation.

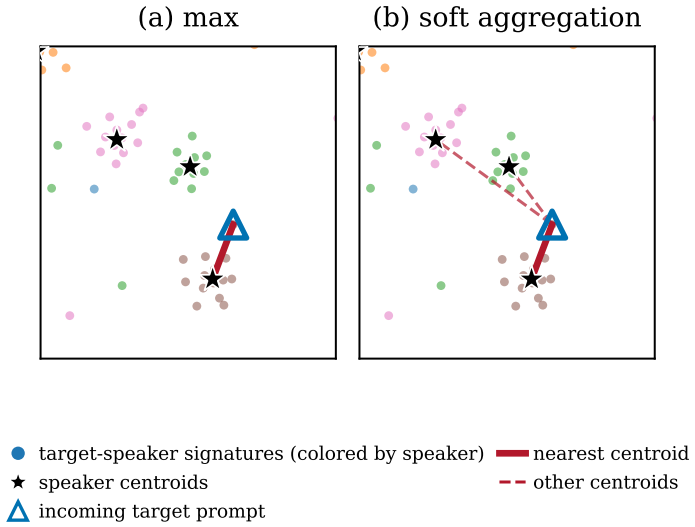

Figure 8: t-SNE of the enrollment signatures for $K = 10$, colored by speaker; stars mark speaker centroids. The same target prompt (triangle) is scored by its nearest centroid under max (a), while a soft aggregation also weighs other centroids (b); similarities are computed in the original signature space.

protocol, and there $r = 40$ spans 98.0% of their energy against 99.6% at $r = 60$; the recovered directions take the target similarity to 0.184, nearly unchanged at $r = 80$, while the remaining speakers are untouched, since the gate holds them at $\gamma \approx 0$ and the projection rank is then irrelevant to them. Under the balanced protocol at $K = 30$, where $r = 40$ already spans 99.3%, the same increase moves neither side: 0.174 and 0.642 at $r = 60$ and at $r = 80$ alike. We keep $r = 40$ everywhere, fixed once at the first-request scale and not tuned per protocol.

Full aggregation grid. We sweep the full plane of soft aggregations $S = \mathrm{LSE}_T$ over the top-N centroid similarities, at the final configuration. Figure 9 reports leak at a matched false-positive rate of 2.5%.

At $K = 10$, two soft-aggregation settings yield leak of 2.0% and 3.0%, versus 4.0% for the maximum, on 200 target prompts. These correspond to four and two fewer missed prompts, respectively. At $K = 100$, evaluated on 2,000 target prompts, the maximum gives the lowest leak in the tested grid; pooling all centroids at high temperature raises leak to 64%. At $K = 10$, the tested multi-centroid variants also increase leak relative to the single-centroid baseline.

Table 8: Eraser rank r on the main protocol (TGU, 3 s prompts) and the balanced protocol at $K = 30$, with the gate held fixed in each column so that only the projection rank changes. Bold marks the setting used throughout.

	Main protocol		Balanced, $K = 30$	
r	SIM-F ↓	SIM-R ↑	SIM-F ↓	SIM-R ↑
20	0.231	0.606	0.209	0.643
40	**0.196**	**0.604**	**0.174**	**0.642**
60	0.184	0.604	0.174	0.642
80	0.183	0.605	0.174	0.642

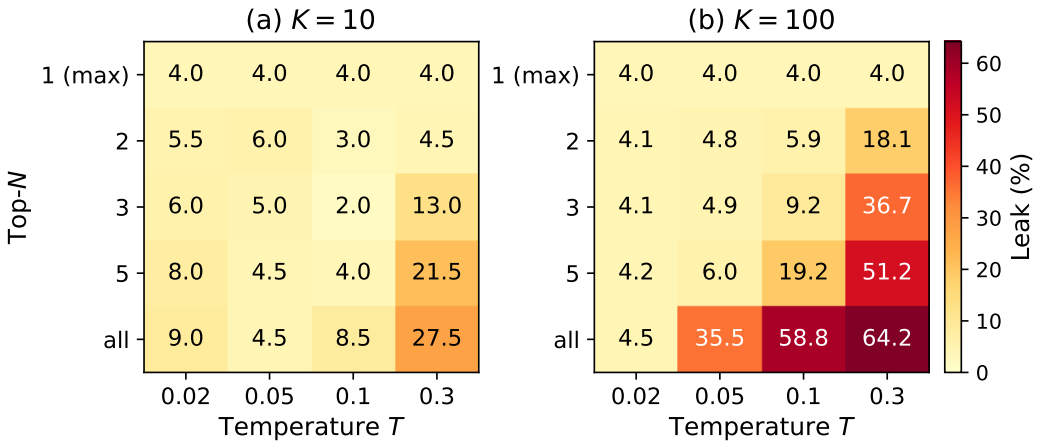

Figure 9: Leak (%) at a matched 2.5% false-positive rate for LSE_T aggregation over the top-N centroids, on a shared colour scale. $N = 1$ is the maximum and is independent of T.

E.1 Projection direction controls

The random control draws 40 orthonormal directions within the 80-dimensional conditioning span, using seed 0. Its gate strengths and 200 evaluation pairs match the top-40 and full-span controls. It captures 49.5% of target feature energy, versus approximately 99.4% for top-40. A second random seed was stopped at 75 pairs and is excluded from the complete-run comparison. These data compare the selected subspace with one random draw, rather than estimating variation across random subspaces.

At fixed rank 40, enrollment-energy coverage is 99.34, 99.36, 99.33, 99.32 and 99.30% for $K = 10, 20, 30, 50, 100$. This explains why the selected rank covers most conditioning energy throughout the evaluated sequence.

Table 9: Projection controls at balanced $K = 10$, with the saved intrinsic gate fixed for all projection rows. SIM-R was not measured for the added random and full-span runs.

Projection	Target energy captured (%)	SIM-F
None (Original)	–	0.631
Random 40 directions	49.5	0.387
Top 40 directions	99.4	0.179
Full conditioning span ($r = 80$)	100.0	0.180

Table 10: Actual generated speech with external ECAPA and intrinsic gating on the same ordered balanced evaluation pairs. Each tier uses $20K$ target and 800 remaining pairs.

	External ECAPA		Intrinsic gate	
K	SIM-F	SIM-R	SIM-F	SIM-R
10	0.165	0.658	0.179	0.647
20	0.169	0.640	0.173	0.636
30	0.169	0.636	0.174	0.642
50	0.180	0.646	0.182	0.648
100	0.196	0.648	0.190	0.648

Table 11: Change in speaker-level SIM-F from enrollment. Only previously enrolled speakers enter each row. Positive values indicate greater output similarity to the target.

K	Prior speakers	Median Δ	10th–90th percentiles	Full range
20	10	-0.000	[-0.043, +0.005]	[-0.058, +0.009]
30	20	+0.001	[-0.026, +0.005]	[-0.058, +0.019]
50	30	+0.001	[-0.008, +0.021]	[-0.065, +0.027]
100	50	+0.002	[-0.003, +0.037]	[-0.066, +0.111]

F External Speaker Verification Gate

External gate construction. The external alternative replaces backbone activation signatures with ECAPA-TDNN embeddings. It computes enrollment whitening with shrinkage 0.1, normalized speaker centroids, and adaptive score normalization against the top 100 synthetic negatives. Its threshold is calibrated on the synthetic bank at $1-\alpha/M = 0.9925$, with $\kappa = 35$. Thus ECAPA controls the same eraser, rather than evaluating a refusal policy. The stored external run uses NumPy's interpolated quantile, while the intrinsic calibration uses the empirical `higher` convention. Internal results are the historical generated outputs described in Appendix B; their soft slopes are not identical to 35 at every tier. These implementation differences are retained rather than treating this as a perfectly matched gate-only ablation. Evaluation pairs and eraser rank are matched, and SIM is measured with WavLM-TDCNN for both alternatives.

ECAPA yields lower SIM-F and higher SIM-R at $K = 10$, while the ordering varies at larger enrollment sizes. At $K = 100$, intrinsic and external gating yield SIM-F of 0.190 and 0.196, respectively, with SIM-R of 0.648 for both. The intrinsic gate supports selective erasing across the evaluated sizes using the backbone's own activation signatures.

G Continual Speaker and Prompt Tracking

G.1 Generated speech for individual speakers

For each enrolled speaker, we follow the same 20 source/target evaluation pairs at every subsequent request. Figure 10 displays the 50 speakers enrolled before the final request; the 50 newcomers at $K = 100$ have no later observation. Each cell averages generated-speech SIM-F over the speaker's pairs. Table 11 summarizes changes relative to each speaker's own enrollment stage, including every speaker with at least one subsequent request.

At $K = 100$, the median change is $+0.0017$, with the central 80% between -0.0026 and $+0.0367$. Four of the 50 earlier speakers increase by more than 0.05; the full range is $[-0.0658, +0.1114]$. Among the first ten speakers, speaker 789 changes from 0.2140 to 0.2677, while speaker 3912 changes from 0.2594 to 0.1936. Thus the group averages in Table 2 are accompanied by individual trajectories rather than interpreted as identical behavior for every speaker. These are repeated evaluations of saved generations, not independent generation-seed trials.

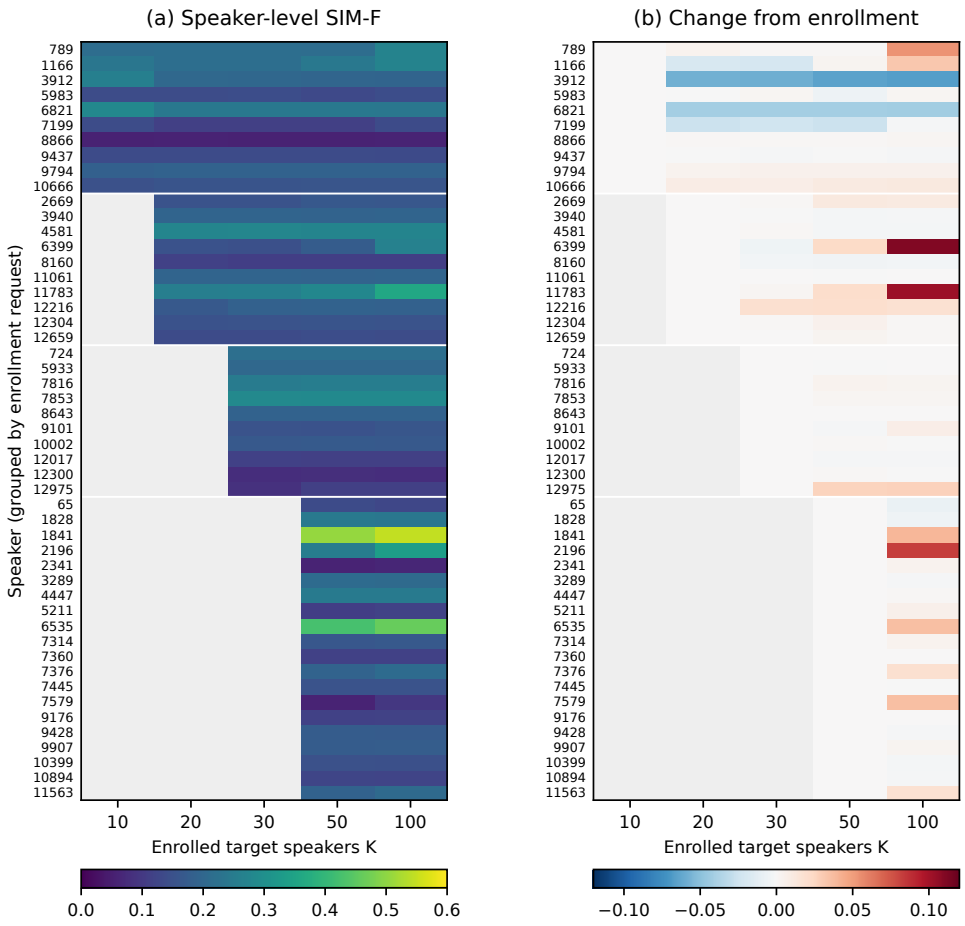

Figure 10: Individual generated-speech trajectories for all 50 speakers enrolled before $K = 100$. Rows are grouped by enrollment request, then speaker ID. (a) Mean SIM-F over the same 20 pairs per speaker. (b) Change from that speaker's enrollment stage. Gray cells precede enrollment.

Table 12: First-request prompt tracking. "Newly below" counts initially above-threshold prompts now below threshold. "Newly above" counts the initially missed prompts now above threshold. Both are relative to $K = 10$, not the preceding row.

K	Leak among original 200 (%)	Newly below / 193	Newly above / 7
10	3.5	0	0
20	0.5	0	6
30	0.5	0	6
50	1.0	1	6
100	3.5	6	6

G.2 Tracking gate decisions on the initial prompts

We follow the same 200 first-request target prompts from $K = 10$ through $K = 100$. Initially 193 score above the gate threshold and seven below it. Table 12 records later threshold crossings among these 193, alongside the overall leak of the original 200 prompts. This is a score-based analysis; a crossing does not by itself quantify recovered output identity under the soft gate.

At $K = 100$, six initially detected prompts fall below threshold, while six initially missed prompts move above it. The unchanged aggregate leak therefore masks individual changes. The six new misses are distributed over speakers 1166, 7199 and 789; the largest within-speaker fraction is 15%

Table 13: Per-speaker target results under the main protocol: similarity of the generated speech to the prompt, and mean applied erasure strength. Ordered by residual similarity.

Speaker	SIM-F	mean γ
3912	0.210	0.94
1166	0.187	0.96
6821	0.183	1.00
9794	0.180	1.00
9437	0.134	1.00
5983	0.123	1.00
789	0.119	1.00
7199	0.104	0.95
8866	0.026	1.00

for speaker 789. Their initial score margins have median 0.991, compared with 1.831 among all initially detected prompts. The eraser also changes: the smallest principal-angle cosine between the $K = 10$ and $K = 100$ subspaces is 0.4457, with 35 of the 40 cosines above 0.99. State-composition equivalence and retention of individual behavior are consequently distinct properties.

H Per-Speaker and Target-Set Identity Analysis

Table 13 breaks the main-protocol target results down by speaker. Nine of the ten enrolled speakers contribute evaluation pairs, in highly unequal numbers; speaker 3912 alone supplies 62% and, having the highest residual similarity, dominates the micro average.

Residual identity and detection. Restricting the comparison to detected prompts still gives speaker-level residuals between 0.026 and 0.183. For speaker 3912, the gate assigns $\gamma > 0.9$ on 93% of prompts and the selected subspace captures 99.3% of conditioning energy, within the observed 99.0–99.5% range. These measurements separate residual similarity after intervention from missed gate decisions. The aggregate gap to TGU is decomposed in Appendix I.

Concentration of missed prompts. The main protocol's 63 misses comprise $46/741$ pairs from speaker 3912, $16/355$ from speaker 1166, and $1/22$ from speaker 7199; the other evaluated target speakers have no missed pairs in this sample. The concentration partly reflects the unequal number of evaluation pairs. The 6.2% rate for speaker 3912 describes observed suppression failures rather than a negligible difference in the overall mean.

H.1 Similarity to the entire enrolled target set

Prompt similarity tests the requested identity; this analysis also checks similarity to other enrolled identities. We form each speaker reference by averaging WavLM-TDCNN embeddings from 20 enrollment recordings and normalizing the mean. For each target output, we measure similarity to its own speaker reference, the maximum over the other enrolled references, and the maximum over the entire enrolled set (FSSIM). The same references score the 800 remaining outputs at each tier and 400 prompt-free generations. Audio is cropped to at most 15 s in this analysis. These reference-based scores therefore have a different comparison unit from the prompt-based SIM in the main tables.

FSSIM stays close to the remaining-output reference level across the evaluated tiers. Maxima increase with the number of candidate identities, making the same-K reference distributions essential. As a genuine-identity comparison, 500 recordings excluded from the reference means score 0.877 on average against their own speaker references (5th–95th percentiles 0.766–0.946). This analysis characterizes generated identities and their tails; it is not a test of information deletion from the backbone.

Table 14: Similarity against the enrolled target set. Own and Other refer to the source speaker's reference and the maximum over other target references. Remaining and Prompt-free are maxima over the same full target set. The last column is the fraction of target outputs whose Other score exceeds the 95th percentile of the remaining-output reference distribution. Each tier evaluates $20K$ target outputs.

K	Own	Other	FSSIM	Remaining	Prompt-free	Exceedance (%)
20	0.185	0.314	0.328	0.332	0.318	5.0
30	0.189	0.339	0.351	0.351	0.338	2.5
50	0.193	0.365	0.378	0.372	0.366	4.4
100	0.200	0.402	0.412	0.415	0.395	2.9

Table 15: Further measurements on the main-protocol generations. ECAPA re-scores the same audio with a different speaker verifier; UTMOS is an automatic mean-opinion-score predictor; spk-ZRF measures closeness to prompt-free generation. Original is the backbone before intervention.

	ECAPA		UTMOS		spk-ZRF
Methods	F ↓	R ↑	F	R	F ↑
Original	0.636	0.652	3.63	4.03	0.840
TruS	0.613	0.619	3.63	3.91	0.841
TGU	**0.080**	0.629	3.98	4.11	**0.859**
SGU	0.100	0.465	4.05	4.10	0.838
ESIAN (Ours)	0.119	**0.640**	4.13	4.03	0.839

I Evaluation Robustness

The main comparison rests on one similarity metric computed by one verifier. This section asks what the conclusions depend on. Table 15 reports three further measurements on the same main-protocol generations.

A second verifier. Re-scoring the main-protocol generations with ECAPA-TDNN (Desplanques et al., 2020) gives SIM-F 0.119 for ESIAN versus 0.636 for Original and SIM-R 0.640 versus 0.652 (Table 15). This corroborates output-level suppression under a second embedding model. TGU and SGU change order on SIM-F under this metric (0.080 versus 0.100), illustrating metric dependence. These are post-generation evaluations; ECAPA is not used by the intrinsic gate in these runs.

Audio quality. The UTMOS results in Table 15 are automatic quality predictions, not human listening judgments. ESIAN and Original both score 4.03 on remaining outputs. Target-side scores are 4.13 for ESIAN and 3.63 for Original; TGU and SGU also score higher than Original on this side. These measurements support preservation of predicted quality, without identifying a unique cause of the increase.

spk-ZRF. Kim et al. (2025) propose spk-ZRF, one minus the Jensen–Shannon divergence between the speaker embeddings of a method's output and those of the same model generating without a prompt; a higher value means the output is closer to a speaker-free voice. On this metric our result is indistinguishable from the Original (0.839 against 0.840), as is the oracle-gated variant (0.840), including when target membership is provided by the oracle gate. The metric rewards a different behavior from the one we implement: it measures how far a method randomises identity, whereas objective (i) concerns suppressing reproduction of the target voice, evaluated through SIM-F. The distinction is visible in the numbers of Kim et al. (2025) themselves, where exact unlearning, retraining from scratch without the target speakers, also scores 0.846, exactly the Original's value, and their KL baseline scores 0.810, below the Original, while reaching SIM-F 0.179. Only methods that finetune the weights to push target prompts toward other voices raise it.

Confidence intervals. The previously computed marginal bootstrap intervals use 10,000 resamples and seed 0. Speaker-resampled macro SIM-F is 0.141 $[0.105, 0.173]$ for ESIAN, 0.148

$[0.105, 0.188]$ for TGU, and 0.148 $[0.112, 0.183]$ for SGU. To compare preservation while accounting for repeated pairs from the same speaker, we additionally align outputs by source/target pair, average each method's differences within each of the 40 remaining speakers, and resample these speaker means with replacement (10,000 draws, NumPy `default_rng(0)`). ESIAN minus TGU gives $+0.0127$ $[+0.0006, +0.0211]$, ESIAN minus TruS gives $+0.0127$ $[-0.0009, +0.0234]$, and ESIAN minus SGU gives $+0.1521$ $[+0.1140, +0.1925]$. Thus the small preservation advantage is not separated from zero against TruS at the speaker level. These intervals quantify evaluation-sample uncertainty for the stored runs; they do not cover generation-seed or enrollment-set variation.

What the micro gap is made of. Among the $1{,}130$ target pairs above threshold, ESIAN has mean SIM 0.1714 and TGU 0.1713 on the same pairs. The other 63 ESIAN pairs have mean SIM 0.6408. Their contribution accounts for approximately 0.0248 of the aggregate 0.0251 SIM-F gap to TGU.

J Erasing Speakers Absent from Training

Setup. The ten speakers are drawn from LibriTTS (Zen et al., 2019) `train-clean-100`, restricted to speakers that do not appear in the LibriHeavy split the backbone was trained on; the `test-clean` split is excluded because it shares its speakers with our remaining set. A speaker is eligible when it also has at least 80 utterances of four seconds or more with a forced alignment, and ten of the eligible speakers are drawn with seed 0. Each speaker supplies 40 enrollment utterances and 20 evaluation pairs disjoint from them. Every protocol in this paper drops a target utterance whose forced alignment leaves more than 10% of its duration unaligned, and that rule removes 30 of the 200 pairs here, leaving 170; LibriTTS alignments trip this filter, while the released LibriHeavy list removes nothing. The run follows the balanced protocol, so the comparable in-domain figure is the 0.179 of Table 2 at $K = 10$ rather than the main-protocol 0.196. Nothing in the method changes: the eraser and the gate are rebuilt from these speakers' enrollment alone, and the threshold is self-calibrated on the same synthetic bank, with no access to remaining speech. Preservation is measured on the same LibriSpeech remaining speakers as before, so the domain shift applies to the target side only.

Results. Target similarity falls from 0.672 to 0.169, and the remaining speakers hold 0.651 against the 0.658 they reach with the gate held shut, a preservation loss of 1.1%. Word error rate moves from 3.9% to 4.4% on the target prompts and stands at 4.4% on the remaining set. The gate separates the two populations at AUC 0.9996, and its self-calibrated threshold lands at 1.4% false positives and no leak. Kim et al. (2025) run their out-of-domain test on a single LibriTTS speaker with a finetuning pass; ours covers ten and trains nothing.