PRIVACY-PRESERVING CLINICAL SPEECH EMBEDDINGS UNDER MULTI-RECORDING IDENTITY-LINKAGE ATTACKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Privacy in learned representations can fail under repeated observation. An embedding that appears privacy-preserving on its own may reveal identity when information from multiple recordings of the same person is combined. We study this problem in clinical speech and introduce MoIRAGE (Multi-observation Identity-Resistant Adversarially Generated Embeddings), a framework for privacy-preserving representation learning under repeated observations. MoIRAGE separates an input embedding into clinically relevant content and an identity-related component, discards the original identity-related component, and replaces it with a synthetic alternative before generating the released embedding. The model is trained against identity-linkage attacks that jointly analyze multiple recordings, while clinical-consistency and diversity objectives preserve clinically relevant information and prevent representation collapse. At inference time, MoIRAGE does not require the participant's diagnosis label. We evaluate MoIRAGE on the UNMC clinical speech dataset using pooled held-out predictions from five participant-disjoint outer folds covering 137 participants, including 127 with clinical labels. MoIRAGE retains performance on the target clinical task, achieving an AUROC of 0.643 compared with 0.567 for the original embeddings. With a single recording ($K = 1$), the strongest identity-linkage $\mathrm{AUC}^{\star}$ decreases from 0.778 to 0.528, close to chance. With 20 recordings ($K = 20$), the strongest identity-linkage $\mathrm{AUC}^{\star}$ decreases from 0.963 for the original embeddings to 0.811 with MoIRAGE, while TPR at 1% FPR decreases from 0.474 to 0.036. Ablation experiments identify replacement of the identity-related component as the central mechanism underlying the privacy reduction, while a teacher-free variant shows similar behavior on Bridge2AI without speaker-teacher supervision. These results show that privacy cannot be adequately evaluated from single observations alone: identity leakage can increase as repeated measurements become available, and MoIRAGE explicitly addresses this multi-observation threat model while retaining performance on the target clinical task.

1 INTRODUCTION

Speech is one of the richest human signals. We can learn many patterns from speech, not only from what is being said, but also from how a person speaks. Beyond the transcript, speech carries information about identity, speaking style, prosody, and health-related characteristics. Modern speech models can capture many of these patterns in compact embeddings, making them useful for a wide range of downstream applications (Chen et al., 2022). In healthcare, these embeddings are increasingly studied for tasks such as disease detection, cognitive assessment, and longitudinal monitoring (Berisha & Liss, 2024). Their ability to preserve clinically useful information makes them attractive for analysis and data sharing, but it also creates an important privacy challenge.

Replacing raw speech with an embedding does not automatically hide the speaker's identity. Identity-related information can still remain in the embedding, allowing an attacker to determine whether different recordings belong to the same participant. Privacy-preserving methods try to reduce this identity information while keeping the representation useful for downstream tasks (Tomashenko et al., 2024; Tayebi Arasteh et al., 2024). This problem becomes more important in clinical speech datasets, where the same participant may contribute multiple tasks, sessions, or longitudinal visits. An attacker can therefore combine several released embeddings from the same person and accumulate identity information across recordings.

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.

We observe this directly in our held-out UNMC naturalistic in-vehicle speech cohort: identity-linkage AUC* increases from 0.778 with one recording ($K = 1$) to 0.963 with 20 recordings ($K = 20$). This shows that privacy measured from a single embedding can underestimate the risk when multiple recordings from the same participant are available.

A useful private representation therefore has to do two things at the same time: preserve clinically relevant information and reduce the identity information that links it back to the participant. Simple perturbation is not enough: Gaussian noise can reduce linkage while also reducing clinical utility, and releasing a learned content representation alone can remain highly linkable when multiple recordings are combined. These observations motivate changing the identity-related information before the embedding is released.

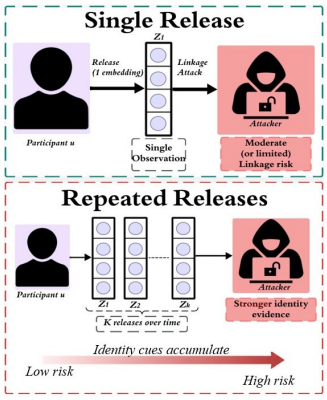


Figure 1: **Privacy risk under repeated observations.** Identity evidence accumulates when multiple recordings from the same participant are available.

We introduce **MoIRAGE** (Multi-observation Identity-Resistant Adversarially Generated Embeddings), a framework for generating clinical speech embeddings with reduced participant linkability. MoIRAGE separates clinically useful content from identity-related information, discards the source identity-related component, and replaces it with a synthetic sample before generating the released embedding. It is also trained against an attacker that combines multiple released embeddings. Clinical-consistency and anti-collapse objectives preserve useful information, and the participant's diagnosis label is not required at inference time. On UNMC, MoIRAGE retains performance on the target clinical task, achieving an AUROC of 0.643 compared with 0.567 for the original embeddings, while reducing identity-linkage AUC* from 0.778 to 0.528 at $K = 1$. Linkage remains substantially reduced as additional recordings are combined, reaching 0.730 at $K = 10$. At $K = 20$, high-confidence linkage also decreases substantially: TPR at 1% FPR falls from 0.474 to 0.036. Removing synthetic identity replacement increases $K = 20$ linkage to 0.956, close to the original representation. A teacher-free variant shows the same qualitative behavior on Bridge2AI-Voice (Bensoussan et al., 2025) development folds.

Contributions. Our contributions are: **(1) Repeated-observation privacy evaluation.** We evaluate participant linkage when collections of released clinical speech embeddings are available, using increasing observation counts, strong linkage attackers, strict operating points, and participant retrieval, and show that single-observation evaluation can underestimate multi-recording privacy risk. **(2) Multi-observation identity-resistant release.** We introduce MoIRAGE, which combines synthetic replacement of the source identity-related component with collection-level adversarial training to reduce participant linkage while retaining performance on the target clinical task. **(3) Mechanistic and empirical analysis.** Targeted ablations identify synthetic identity replacement as the central mechanism underlying the privacy reduction; repeated-release variation decreases approximately as $1/R$ and a teacher-free Bridge2AI development replication shows similar qualitative privacy-utility behavior in a different cohort and feature space.

2 Related Work

Speech anonymization. Speech anonymization aims to reduce speaker identity while preserving useful linguistic and paralinguistic information. The VoicePrivacy benchmarks formalized this privacy-utility setting using speaker-verification attacks (Tomashenko et al., 2022). Later methods transform or replace speaker representations while trying to preserve speech quality and content (Miao et al., 2023). This trade-off is especially important in clinical speech, where disease-related information must remain useful after anonymization. Recent work shows that pathological speech can be anonymized while retaining diagnostic information, although the trade-off varies across clinical conditions (Tayebi Arasteh et al., 2024). Woszczyk et al. (2024) study privacy-preserving speaker embeddings for dementia detection using prosody-driven disentanglement. In contrast, we focus on participant linkage when embeddings from multiple recordings are jointly available and train the release mechanism explicitly for this multi-observation setting.

Related voice-anonymization methods also replace source speaker attributes with pseudo-speaker representations. Chen et al. model pseudo-speaker attributes as a learned distribution and sample a pseudo-speaker representation for anonymization (Chen et al., 2025). This is closely related to the replacement component of MoIRAGE. Our setting differs in that the released object is a learned

clinical embedding rather than a resynthesized waveform, and the release mechanism is optimized directly against participant linkage across multiple observations while preserving downstream clinical information.

Beyond speech, representation learning has studied how to preserve information needed for a task while reducing information about unwanted or sensitive factors. Adversarial methods make a target task predictable while making a nuisance attribute difficult to recover (Xie et al., 2017; Feutry et al., 2018). Information-theoretic methods similarly reduce dependence on nuisance factors (Moyer et al., 2018; Alemi et al., 2017). These approaches motivate the content-identity separation used in MoIRAGE. Our approach differs in what is released: instead of only suppressing the unwanted factor, MoIRAGE replaces the source identity-related component with a synthetic alternative before generating the released embedding.

Privacy-preserving speech representations. Adversarial representation learning can reduce speaker classification, but this does not necessarily prevent identity recovery by a different verification attacker (Srivastava et al., 2019). Other work transforms pretrained speech embeddings toward alternative speaker identities while preserving downstream tasks such as emotion, depression, and intent recognition (Tran & Soleymani, 2023). These findings motivate evaluating privacy with attackers that differ from the model used during training. MoIRAGE further trains against attackers that combine information across several released embeddings.

Multiple recordings can provide stronger identity evidence than a single recording. Multi-enrollment speaker-verification systems combine several utterances to improve identity matching (Zeng et al., 2022). More recently, Garg et al. show that aggregating multiple anonymized utterances improves speaker verification, demonstrating that speaker-discriminative information can accumulate even after anonymization (Garg et al., 2026). More generally, privacy can weaken when information from repeated releases is combined (Dwork & Roth, 2014).

We study this concern for released clinical speech embeddings by measuring participant linkage as the number of available recordings increases and by training the release mechanism directly against collection-level linkage. Existing work has considered speech anonymization, pseudo-speaker replacement, invariant representations, and multi-utterance identity matching, but these directions are typically studied separately. MoIRAGE combines synthetic identity-related replacement with multi-observation adversarial training and repeated-observation privacy evaluation in a clinical representation-learning setting.

3 Privacy Setting

We consider a setting where the same participant can contribute multiple speech recordings. Let $x_i^{(j)}$ denote recording j from participant i, and let the release mechanism M produce $z_i^{(j)} = M(x_i^{(j)})$. For K available recordings, the attacker observes $\mathcal{Z}_i^K = \{z_i^{(1)}, \ldots, z_i^{(K)}\}$. The attacker's goal is to determine whether two collections of released embeddings belong to the same participant.

Threat model. We assume that the attacker has access to the released embeddings and may know how the release mechanism is constructed. Privacy therefore does not depend on keeping MoIRAGE secret. The attacker does not need access to the target participant's raw audio or clinical diagnosis, and we assume that the attacker knows the observation count K.

Attack models are trained and selected using participants separate from the final held-out evaluation participants. Same-participant comparisons use different enrollment and probe recordings, so the two sides of a linkage trial do not reuse the same recording. For each K, the attacker combines the available embeddings using similarity measures, summary statistics, or learned set models and predicts whether two collections belong to the same participant.

We report linkage mainly using orientation-invariant AUC:

$$\mathrm{AUC}^\star = \max\left(\mathrm{AUC}, 1 - \mathrm{AUC}\right). \tag{1}$$

A value of 0.5 represents chance-level linkage, while values closer to 1 indicate stronger linkage. We use $\mathrm{AUC}^\star$ because a consistently inverted linkage score can be reversed without losing the identity information it contains.

We do not claim that the released embeddings are anonymous. We measure how much participant linkage remains and how that risk changes as K increases.

4 MoIRAGE

We introduce **MoIRAGE** (Multi-observation Identity-Resistant Adversarially Generated Embeddings), a framework for generating clinical speech embeddings with reduced participant linkability.

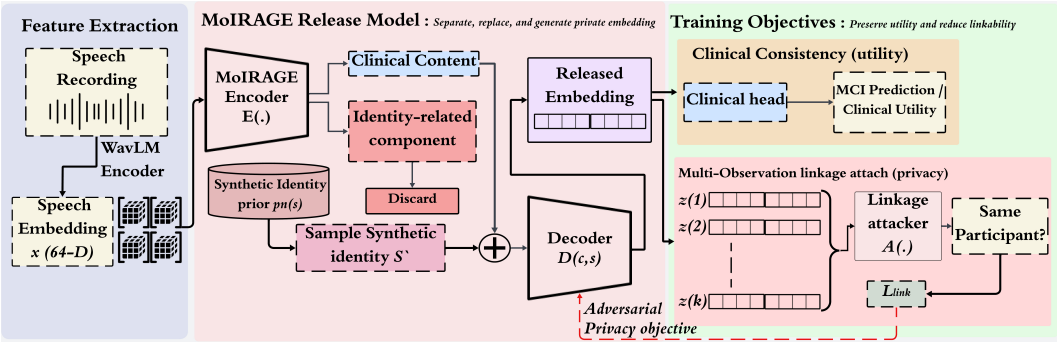


Figure 2: **Overview of MoIRAGE.** WavLM converts each speech recording into an input embedding x, which the MoIRAGE encoder separates into clinically useful content c and a source identity-related component s_{src}. The source component is discarded and replaced with a synthetic sample $\tilde{s}$ before the decoder generates the released embedding z. Clinical objectives preserve downstream utility, while a multi-observation attacker operates on sets of K released embeddings and provides the adversarial signal used to reduce participant linkability.

The main idea is to preserve clinically useful information while changing the identity-related information used to generate the released embedding.
Given an input embedding x, the encoder separates it into

$$E(x) = (c, s_{\text{src}}), \tag{2}$$

where c represents clinically useful content and s_{src} represents identity-related variation from the source recording.
Instead of reusing s_{src}, MoIRAGE samples a synthetic identity-related component from a learned class-conditional distribution:

$$p_\eta(s \mid y') = \mathcal{N}\left(m_{y'}, \operatorname{diag}(e^{v_{y'}})\right), \qquad \tilde{s} \sim p_\eta(s \mid y'). \tag{3}$$

The released embedding is then

$$z = \operatorname{norm}\left(D(c, \tilde{s})\right), \tag{4}$$

where D is the decoder and $\operatorname{norm}(\cdot)$ denotes ℓ_2 normalization. A new $\tilde{s}$ can be sampled for each release. In the label-blind setting, y' is sampled from the training class distribution rather than taken from the participant's diagnosis.

Content-identity factorization. The content branch is trained to retain information needed for the clinical task, while the identity branch captures additional participant-related variation. We apply a variational bottleneck to c (Alemi et al., 2017), reconstruct the original embedding while stopping reconstruction gradients through c, and penalize redundancy between c and s_{src}. These constraints reduce the pressure for the content representation to store all information from the input. Full loss definitions are provided in the appendix.

Synthetic identity replacement. Factorization alone does not guarantee low participant linkage because the content representation can still contain identity information. MoIRAGE therefore discards s_{src} and replaces it with a synthetic sample $\tilde{s}$ before generating the released embedding. The clinical content still comes from the source recording, but the original identity-related component is not directly reused.

Clinical preservation. Replacing identity-related information is useful only if the clinical signal remains stable. We train clinical heads on both the content representation and the released embedding. For each input, two independent releases are generated and encouraged to produce similar clinical predictions. We also distill the clinical heads toward out-of-fold logistic-regression probabilities. These objectives help preserve clinical information as the synthetic identity-related component changes. Full definitions are given in the appendix.

Multi-observation adversarial training. MoIRAGE is trained against an attacker that compares collections of released embeddings. We use a permutation-invariant DeepSets-style attacker (Zaheer et al., 2017). For a set A,

$$e(A) = \rho_\psi\left([\operatorname{mean} \phi_\psi(A) \parallel \operatorname{std} \phi_\psi(A)]\right), \tag{5}$$

and two sets A and B are compared using

$$u(A, B) = \mathrm{MLP}\left([|e(A) - e(B)| \parallel e(A) \odot e(B)]\right). \tag{6}$$

The comparison is symmetric and does not depend on the order of the embeddings. The attacker tries to recover participant linkage, while MoIRAGE is trained to make this comparison difficult.

Speaker supervision and teacher-free training. For the primary UNMC model, the identity-related representation is supervised using a frozen ECAPA speaker embedding. ECAPA supervises only the identity branch and is used only during training. The teacher-free variant removes this supervision while keeping the remaining MoIRAGE objectives unchanged. We use this variant for the Bridge2AI replication.

Preventing representation collapse. A model could reduce linkage by mapping many inputs to nearly the same released embedding. We therefore regularize the released space using covariance, eigenvalue, effective-rank, and class-conditional mean constraints. Checkpoint selection also requires minimum covariance-trace and effective-rank values. Full definitions and diagnostics are provided in the appendix.

Overall objective. The complete training objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{clinical}} + \mathcal{L}_{\mathrm{factorization}} + \mathcal{L}_{\mathrm{multi}} + \mathcal{L}_{\mathrm{div}} + \lambda_{\mathrm{spk}}\mathcal{L}_{\mathrm{spk}} + \mathcal{L}_{\mathrm{dual}}. \tag{7}$$

Here, $\mathcal{L}_{\mathrm{clinical}}$ preserves clinical information, $\mathcal{L}_{\mathrm{factorization}}$ separates content and identity-related variation, $\mathcal{L}_{\mathrm{multi}}$ reduces collection-level linkage, and $\mathcal{L}_{\mathrm{div}}$ prevents representation collapse. $\mathcal{L}_{\mathrm{dual}}$ enforces the clinical utility constraint. Exact weights and loss definitions are provided in the appendix.

MoIRAGE is designed to reduce empirical participant linkability under repeated observations while preserving information needed for downstream clinical analysis. It provides an empirical privacy mechanism rather than a formal differential-privacy guarantee. The progression from simpler baselines and earlier design variants to the final MoIRAGE architecture is summarized in Appendix B.

5 Experimental Setup

Datasets and splits. We evaluate MoIRAGE on two clinical speech datasets. UNMC is our primary dataset for development and held-out evaluation. The cohort contains 137 participants, 3,869 participant-days, and 10,286 participant-speech chunks of up to 60 seconds. Of the 137 participants, 127 have the binary MCI/control label. Participants contribute a median of 50 recordings (IQR 29-96), providing repeated observations from the same person.

We use five participant-disjoint outer folds. The TEST partitions are disjoint and together contain all 137 participants, so every participant is evaluated once by a model that did not see that participant during training. Each recording is represented using WavLM (Chen et al., 2022). Scaling and the 64-dimensional PCA projection are fitted on TRAIN participants only and then applied unchanged to validation and TEST data. The architecture, objectives, hyperparameters, checkpoint-selection criteria, and attacker-selection procedure were established using TRAIN and validation data before evaluation of the corresponding outer TEST partitions; no TEST participant was used for release-model or attacker selection. Full cohort provenance, split statistics, and preprocessing details are provided in Appendix E.

We also evaluate the teacher-free MoIRAGE variant on Bridge2AI-Voice (Bensoussan et al., 2025). The prepared development cohort contains 625 participants, of whom 153 have the binary clinical label used for clinical-utility evaluation. Bridge2AI uses five independently generated development splits, each with 500 training and 125 validation participants; the validation sets are therefore not mutually exclusive across folds. Across the five validation folds, 155 labeled prediction rows correspond to 113 unique participants, including 29 positive cases. For the pooled participant-level clinical evaluation, predictions from repeated validation appearances of the same participant are averaged before computing AUROC. A separate locked partition of 157 participants was not available for the reported experiments, so Bridge2AI results are treated as an external development replication rather than a second confirmatory test.

Clinical utility. We evaluate performance on the target clinical task after the embedding is transformed. We use logistic regression, Elastic Net, and a linear SVM. Logistic-regression AUROC is our primary utility metric, with AUPRC and balanced accuracy reported as additional measures. All primary results use the label-blind MoIRAGE release setting.

Privacy evaluation. We measure whether released embeddings from different recordings can still be linked to the same participant. Enrollment and probe recordings come from different participant-days. For each observation count K, we compare collections of recordings from the enrollment and

Table 1: **Main UNMC test results.** Lower linkage is better; 0.5 represents chance-level linkage.

Metric	Original	MoIRAGE
Clinical AUROC	0.567	**0.643**
Linkage AUC*, $K = 1$	0.778	**0.528**
Linkage AUC*, $K = 20$	0.963	**0.811**
TPR @ 1% FPR, $K = 20$	0.474	**0.036**
Rank-1, $K = 20$	0.716	**0.161**

probe pools. We mainly evaluate $K \in \{1, 2, 5, 10, 20\}$. For same-participant trials, if fewer than $2K$ eligible recordings are available, enrollment and probe use the largest equal-size disjoint sets that can be formed without reusing a recording; the effective set size can therefore be smaller than K.

The pooled UNMC linkage evaluation contains 3,755 comparisons per K: 137 same-participant and 3,618 different-participant comparisons. Learned attackers are trained using balanced batches of same-participant and different-participant set pairs. Attacker-training participants are separate from the final TEST participants.

We evaluate several attacker families, including cosine templates, statistical aggregation with an MLP, DeepSets (Zaheer et al., 2017), attention-based set models, and adaptive cross-space attacks. Attacker architecture and capacity are selected using validation participants and then fixed before TEST evaluation. Since privacy estimates depend on attacker strength, we report the strongest observed attacker at each K.

Baselines and statistical evaluation. We compare MoIRAGE with the original embeddings, Gaussian perturbation, a content-only representation, and Iterative Nullspace Projection (INLP) (Ravfogel et al., 2020), an established linear concept-erasure method. We also remove synthetic identity replacement, multi-observation adversarial training, and ECAPA speaker supervision to measure the contribution of each component.

Our primary privacy metric is $\mathrm{AUC}^\star$. We additionally report EER, TPR at 1% and 5% FPR, and participant retrieval. Clinical confidence intervals use 1,000 participant-level bootstrap resamples. Privacy intervals for the learned set attackers use 500 resamples, while the paired clinical AUROC difference uses 2,000 resamples. We report 95% percentile intervals. Full evaluation and implementation details are provided in the appendix.

6 Results

MoIRAGE reduces participant linkage while retaining performance on the target clinical task. Table 1 summarizes the main held-out UNMC results. Linkage is substantially reduced at both single and repeated observations, although repeated recordings still increase identity risk.

6.1 Privacy under repeated observations

Figure 3(a) shows how participant linkage changes as more recordings become available. For the original embeddings, linkage increases from 0.778 at $K = 1$ to 0.963 at $K = 20$. MoIRAGE reduces linkage to 0.528 at $K = 1$ and 0.811 at $K = 20$, and remains lower across all evaluated observation counts. At $K = 50$, linkage is 0.965 for the original embeddings and 0.854 for MoIRAGE.

These results show both the benefit and the limitation of MoIRAGE. The method slows the accumulation of identity information, but repeated recordings still increase participant linkage. Privacy is therefore improved, not solved.

Why does linkage increase with repeated observations? A simple random-effects view helps explain this behavior. Let a released representation be approximated as

$$z_{ijr} = \mu_i + \delta_{ij} + \epsilon_{ijr},$$

where μ_i is participant-stable information, δ_{ij} is recording-specific variation, and ϵ_{ijr} is stochastic variation introduced by synthetic replacement. As more recordings from the same participant become available, the participant-stable component remains, while recording-specific and stochastic variation can average out.

To isolate the stochastic component introduced by synthetic replacement, we generate multiple independent releases of the same recording and average R releases. Figure 4(a) shows that replacement-induced dispersion decreases approximately as $R^{-1.03}$, closely matching the expected $1/R$ scaling.

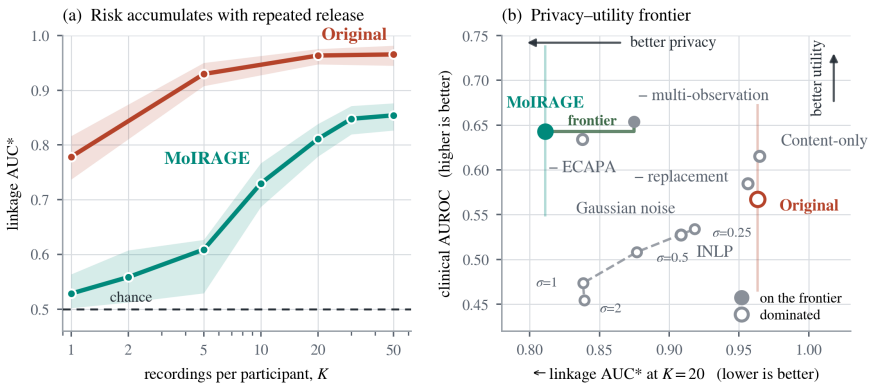
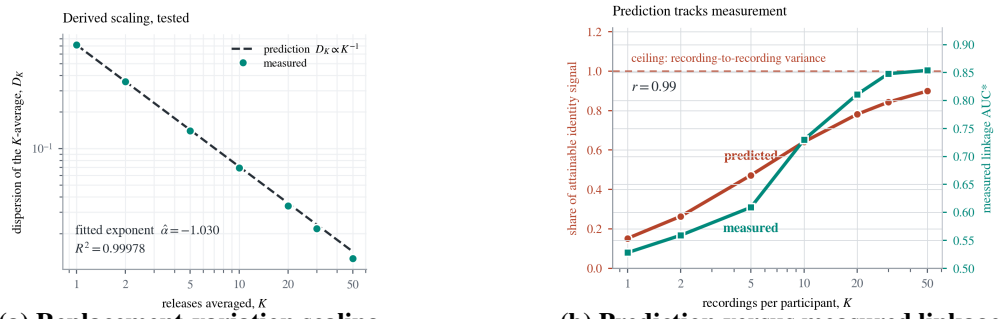


Figure 3: **Repeated-observation privacy and privacy-utility trade-off on UNMC. (a)** Participant linkage increases as more recordings from the same participant become available, while MoIRAGE remains below the original embeddings across observation counts. **(b)** Privacy-utility comparison at $K = 20$, including Gaussian perturbation at $\sigma \in \{0.25, 0.5, 1, 2\}$ and INLP; lower linkage and higher clinical AUROC are preferred.



(a) Replacement-variation scaling **(b) Prediction versus measured linkage**

Figure 4: **Mechanistic analysis of residual linkage under repeated observations. (a)** When independent synthetic releases of the same recording are averaged, replacement-induced dispersion closely follows the predicted $D_R \propto R^{-1}$ scaling, with fitted exponent $\hat{\alpha} = -1.03$. **(b)** A variance-based prediction of attainable identity signal tracks the measured increase in linkage $\mathrm{AUC}^{\star}$ as the attacker receives more recordings ($r = 0.99$).

Here, R denotes independent synthetic releases of the same recording and is distinct from K, which denotes different recordings available to the linkage attacker.

Across held-out folds, the variance decomposition attributes approximately 82.9% of the observed variation to the synthetic replacement draw, 14.8% to recording-within-participant variation, and 2.3% to participant-stable variation. Although the stable component is small for an individual release, it is not removed by averaging independent release variation.

Figure 4(b) gives a complementary diagnostic. The variance-based prediction of attainable identity signal rises with the number of recordings available to the attacker and closely tracks measured linkage across K ($r = 0.99$). Together, these analyses are consistent with stable participant information becoming more visible as stochastic and recording-specific variation is averaged across observations. Full derivations and variance-decomposition details are provided in Appendix M.

6.2 Privacy-utility trade-off and mechanism

Figure 3(b) shows the privacy-utility trade-off. Across $\sigma \in \{0.25, 0.5, 1, 2\}$, Gaussian perturbation traces a trade-off between lower linkage and lower target-task performance. The content-only representation remains highly linkable. INLP reduces linkage from 0.778 to 0.651 at $K = 1$, but linkage rises to 0.908 at $K = 20$ while clinical AUROC is 0.528. MoIRAGE reaches lower repeated-observation linkage ($\mathrm{AUC}^{\star} = 0.811$ at $K = 20$) with higher target-task AUROC (0.643).

The 95% participant-bootstrap AUROC intervals are $[0.549, 0.738]$ for MoIRAGE and $[0.465, 0.672]$ for the original embeddings. The paired AUROC difference is $+0.075$, with a 95%

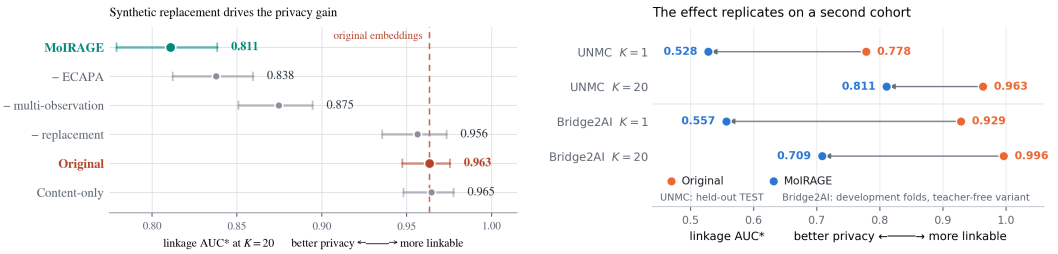


Figure 5: **Mechanism ablation and external development replication.** **(a)** Removing synthetic identity replacement restores $K = 20$ participant linkage close to the original representation; removing multi-observation training or ECAPA supervision produces smaller increases. **(b)** MoIRAGE reduces participant linkage on held-out UNMC and shows the same qualitative pattern in the Bridge2AI development replication.

interval of $[-0.010, 0.163]$. Since this interval includes zero, we interpret the result as retaining performance on the target clinical task rather than as evidence that MoIRAGE improves diagnosis. For MoIRAGE, AUPRC is 0.706 and balanced accuracy is 0.617.

The same pattern appears with other linear classifiers. Elastic Net reaches an AUROC of 0.641 with MoIRAGE and 0.534 with the original embeddings, while linear SVM reaches 0.652 and 0.463. Predictions are also stable across synthetic releases. Over 20 releases of the same input, the average standard deviation of the logistic-regression diagnosis probability is 0.024. Full classifier-specific utility and release-stability results are provided in the appendix.

Figure 5(a) shows that synthetic identity replacement is central to the privacy gain. Keeping the original identity-related component increases $K = 20$ linkage from 0.811 to 0.956, close to the original embedding at 0.963. The content-only representation also remains highly linkable ($\mathrm{AUC}^{\star} = 0.965$ at $K = 20$). Factorization alone is therefore not enough.

Multi-observation training provides an additional benefit. Removing it increases $K = 20$ linkage to 0.875. Removing ECAPA supervision has a smaller effect, increasing linkage to 0.838. At $\sigma = 1$, Gaussian noise reaches 0.838 linkage at $K = 20$, but clinical AUROC falls to 0.473. This shows that low linkage can also be obtained at the cost of target-task performance, while MoIRAGE retains substantially higher target-task performance. Additional geometry diagnostics in the appendix verify that the reduction in linkage is not explained by complete representation collapse.

6.3 High-confidence linkage and attacker strength

Overall AUC does not show how well an attacker performs when only a small number of false matches are allowed. At 1% FPR, attacker TPR decreases from 0.474 to 0.036. At 5% FPR, it decreases from 0.861 to 0.212. Rank-1 participant identification decreases from 0.716 to 0.161. Thus, although linkage remains above chance at $K = 20$, high-confidence identification becomes substantially harder. Full EER, rank-5, and mean-reciprocal-rank results are reported in the appendix.

We also test whether the remaining leakage is caused by a weak attacker. At $K = 20$, a statistical MLP reaches an $\mathrm{AUC}^{\star}$ of 0.811, compared with 0.659 for the DeepSets attacker used during training. Increasing the statistical MLP's width, depth, and training budget gives little additional improvement on validation data. We therefore report the strongest observed attack in the main privacy results. A score-level cosine diagnostic is provided in Appendix Figure 7.

Across the 45 fold-seed runs formed by five held-out folds, three generator seeds, and three attacker seeds at $K = 20$, mean attacker $\mathrm{AUC}^{\star}$ was 0.840 with a between-run standard deviation of 0.039.

6.4 External development replication on Bridge2AI

We next evaluate the teacher-free MoIRAGE variant on Bridge2AI-Voice without changing the architecture. Under the cross-task single-release protocol, linkage decreases from 0.928 for the original embeddings to 0.557 with MoIRAGE. Under the separate all-tasks repeated-observation setting, linkage remains lower as more recordings are combined. At $K = 20$, the DeepSets and statistical MLP attackers reach $\mathrm{AUC}^{\star}$ values of 0.709 and 0.677, respectively, compared with approximately 0.996 for the original embeddings. Figure 5(b) summarizes the linkage reduction across UNMC and the Bridge2AI development replication.

Clinical information is also retained on Bridge2AI. Under the primary label-blind multi-draw protocol, repeated validation predictions are averaged within participant before evaluation. Participant-level AUROC is 0.948 [0.862, 0.996] with MoIRAGE and 0.913 [0.820, 0.974] with the original

embeddings, evaluated on 113 unique held-out participants, including 29 positive cases. Because Bridge2AI does not use a fully untouched TEST partition, we report these results only as an external development replication rather than a second confirmatory test.

7 Discussion and Limitations

Our results show that privacy risk in clinical speech representations should be considered at the participant level rather than one recording at a time. Clinical datasets are often longitudinal or contain multiple tasks from the same participant, allowing identity evidence to accumulate across released embeddings. The increase in linkage from $K = 1$ to larger collection sizes shows that single-recording evaluation can substantially underestimate this risk.

MoIRAGE addresses this setting by replacing the source identity-related component before release and training directly against collection-level linkage attacks. The ablations show that synthetic identity replacement is central to the observed privacy reduction, while multi-observation adversarial training provides an additional benefit as more recordings become available. Importantly, this reduction is achieved while retaining performance on the target clinical task and without requiring the participant's diagnosis label at inference time.

The remaining linkage at large K also reveals an important limitation of representation-level privacy. On UNMC, linkage reaches $\mathrm{AUC}^{\star} = 0.811$ at $K = 20$ and 0.854 at $K = 50$, indicating that residual participant information can still become recoverable when many observations are combined. MoIRAGE should therefore be interpreted as reducing empirical re-identification risk under the evaluated threat model rather than providing a formal anonymity or differential-privacy guarantee.

The current study focuses on representation-level linkage and a limited set of clinical tasks and speech representations. Future work should examine longer longitudinal histories, additional speech encoders and clinical conditions, stronger adaptive attackers, and mechanisms with formal privacy guarantees. The Bridge2AI experiments provide supporting evidence in a different cohort and feature space, although they are reported as an external development replication rather than an independent confirmatory test.

8 Conclusion

We introduced MoIRAGE for reducing participant identity linkage in clinical speech embeddings under repeated observations. Our results show that privacy measured from a single released embedding can give an incomplete picture, as identity evidence becomes substantially stronger when multiple recordings from the same participant are combined. MoIRAGE addresses this setting by separating clinically useful content from identity-related information, replacing the source identity component before release, and training against multi-recording linkage attacks. On UNMC, it reduces single-recording linkage from 0.778 to 0.528 and remains less linkable under repeated observations while retaining target-task performance, targeted ablations further show that synthetic identity replacement is central to this reduction. More broadly, our findings suggest that privacy mechanisms for longitudinal and repeatedly sampled clinical data should be designed and evaluated at the level of repeated observations, rather than one release at a time.

AI Use Statement

Generative AI tools were used to assist with drafting and polishing portions of the manuscript and with retrieval and discovery of relevant literature. All AI-assisted text and references were reviewed and verified by the authors. The authors take responsibility for the final content of the paper.

Ethics Statement

The UNMC data are drawn from an IRB-approved longitudinal study in which participants provided informed consent. Raw speech recordings are not publicly distributed because they contain sensitive participant information. This work evaluates representation-level privacy risk and does not claim anonymity or a formal differential-privacy guarantee. Bridge2AI data are used subject to the access conditions of the dataset maintainers.

References

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *International Conference on Learning Representations*, 2017.

Yael Bensoussan, Alexandros Sigaras, Anais Rameau, Olivier Elemento, Maria Powell, David Dorr, Philip Payne, Vardit Ravitsky, Jean-Christophe Bélisle-Pipon, et al. Bridge2ai-voice: An

ethically-sourced, diverse voice dataset linked to health information. *PhysioNet*, 2025. doi: 10.13026/k81f-qr68. Version 3.0.0, RRID:SCR_007345.

Visar Berisha and Julie M. Liss. Responsible development of clinical speech AI: Bridging the gap between clinical research and technology. *npj Digital Medicine*, 7:208, 2024. doi: 10.1038/s41746-024-01199-1.

Liping Chen, Wenju Gu, Kong Aik Lee, Wu Guo, and Zhen-Hua Ling. Pseudo-speaker distribution learning in voice anonymization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 33:272–285, 2025. doi: 10.1109/TASLP.2024.3519879.

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.

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

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4):211–407, 2014. doi: 10.1561/0400000042.

Clément Feutry, Pablo Piantanida, Yoshua Bengio, and Pierre Duhamel. Learning anonymized representations with adversarial neural networks. *arXiv preprint arXiv:1802.09386*, 2018.

Ashi Garg, Cristina Aggazzotti, Leibny Paola García-Perera, and Nicholas Andrews. Multimodal speaker verification as a threat to speaker anonymization. *arXiv preprint arXiv:2607.19636*, 2026.

Xiaoxiao Miao, Xin Wang, Erica Cooper, Junichi Yamagishi, and Natalia Tomashenko. Speaker anonymization using orthogonal householder neural network. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:3681–3695, 2023. doi: 10.1109/TASLP.2023.3313429.

Daniel Moyer, Shuyang Gao, Rob Brekelmans, Greg Ver Steeg, and Aram Galstyan. Invariant representations without adversarial training. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7237–7256. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.647.

Brij Mohan Lal Srivastava, Aurélien Bellet, Marc Tommasi, and Emmanuel Vincent. Privacy-preserving adversarial representation learning in ASR: Reality or illusion? In *Proceedings of Interspeech*, pp. 3700–3704, 2019. doi: 10.21437/Interspeech.2019-2415.

Soroosh Tayebi Arasteh, Tomás Arias-Vergara, Paula Andrea Pérez-Toro, Tobias Weise, Kai Packhäuser, Maria Schuster, Elmar Noeth, Andreas Maier, and Seung Hee Yang. Addressing challenges in speaker anonymization to maintain utility while ensuring privacy of pathological speech. *Communications Medicine*, 4:182, 2024. doi: 10.1038/s43856-024-00609-5.

Natalia Tomashenko, Xin Wang, Emmanuel Vincent, Jose Patino, Brij Mohan Lal Srivastava, Paul-Gauthier Noé, Andreas Nautsch, Nicholas Evans, Junichi Yamagishi, Benjamin O'Brien, Anaïs Chanclu, Jean-François Bonastre, Massimiliano Todisco, and Mohamed Maouche. The VoicePrivacy 2020 challenge: Results and findings. *Computer Speech & Language*, 74:101362, 2022. doi: 10.1016/j.csl.2022.101362.

Natalia Tomashenko, Xiaoxiao Miao, Pierre Champion, Sarina Meyer, Xin Wang, Emmanuel Vincent, Michele Panariello, Nicholas Evans, Junichi Yamagishi, and Massimiliano Todisco. The VoicePrivacy 2024 challenge evaluation plan. *arXiv preprint arXiv:2404.02677*, 2024.

Minh Tran and Mohammad Soleymani. Privacy-preserving representation learning for speech understanding. In *Proceedings of Interspeech*, pp. 2858–2862, 2023. doi: 10.21437/Interspeech.2023-2138.

Dominika Woszczyk, Ranya Aloufi, and Soteris Demetriou. Prosody-driven privacy-preserving dementia detection. In *Proceedings of Interspeech*, pp. 1760–1764, 2024. doi: 10.21437/Interspeech.2024-2137.

Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, and Graham Neubig. Controllable invariance through adversarial feature learning. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Chang Zeng, Xin Wang, Erica Cooper, Xiaoxiao Miao, and Junichi Yamagishi. Attention back-end for automatic speaker verification with multiple enrollment utterances. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6717–6721, 2022. doi: 10.1109/ICASSP43922.2022.9746688.

A FULL MOIRAGE ARCHITECTURE

Input representation. Each UNMC participant-speech chunk is represented as a 64-dimensional vector $x \in \mathbb{R}^{64}$. We use WavLM-base (Chen et al., 2022). The hidden states from the last four transformer layers are averaged over time, producing four 768-dimensional vectors. These vectors are concatenated into a 3072-dimensional representation and ℓ_2-normalized.Each participant-speech chunk is represented independently. Standardization and PCA are then fitted using TRAIN participants only, reducing each recording-level vector to 64 dimensions. The resulting vector is again ℓ_2-normalized.

Encoder. The encoder first applies two shared blocks,

$$64 \to 128 \to 128,$$

where each block contains a linear layer, GELU activation, and LayerNorm. Four linear heads of size $128 \to 32$ produce the mean and log-variance of two diagonal Gaussian distributions:

$$q(c \mid x) = \mathcal{N}\big(\mu_c, \mathrm{diag}(e^{\ell_c})\big), \qquad q(s \mid x) = \mathcal{N}\big(\mu_s, \mathrm{diag}(e^{\ell_s})\big). \tag{8}$$

Both c and s are 32-dimensional. Log-variances are clamped to $[-6, 3]$.

Content used for release. The released representation uses the posterior mean μ_c rather than a sample from $q(c \mid x)$. Stochasticity in the content branch is used only by the bottleneck objective. The identity-related component is sampled during training and replaced before release.

Learned identity-related prior. The synthetic identity-related component is sampled from a learned class-conditional Gaussian prior:

$$p(s \mid y) = \mathcal{N}(m_y, \mathrm{diag}(e^{v_y})). \tag{9}$$

The two classes each have a learned 32-dimensional mean and log-variance. Both parameter tables are initialized at zero, so training begins from $\mathcal{N}(0, I)$. The log-variance is clamped to $[-6, 3]$. The prior is trained jointly with the rest of MoIRAGE and is frozen with the final model.

Decoder. A single decoder is used for reconstruction and release generation. Its architecture is

$$64 \to 128 \to 128 \to 64,$$

with GELU and LayerNorm after the first two linear layers. Reconstruction uses

$$\hat{x} = D([\mathrm{sg}(\mu_c), s]), \tag{10}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient. The reconstruction is not normalized. The released embedding is

$$z = \frac{D([\mu_c, \tilde{s}])}{\|D([\mu_c, \tilde{s}])\|_2}, \qquad \tilde{s} \sim p(s \mid y'). \tag{11}$$

A fresh $\tilde{s}$ is sampled for each release.

Label-blind release. The primary release setting does not use the participant's diagnosis label. For each released row, the conditioning class y' is sampled from the class distribution observed in TRAIN. The participant's true diagnosis is ignored. An oracle version that uses the true diagnosis is kept only as a reference and is not the primary result.

Training-only components. MoIRAGE uses additional heads during training. A speaker head maps the 32-dimensional identity-related representation to a 192-dimensional ℓ_2-normalized ECAPA-TDNN speaker representation (Desplanques et al., 2020). Two small clinical heads operate on pooled content and released embeddings. A DeepSets linkage attacker (Zaheer et al., 2017) operates on sets of releases. These components are used for training only. The released model requires only the encoder, learned prior, and decoder.

Model size. The full training model contains 104,576 trainable parameters, including the learned prior parameters.

B EVOLUTION OF MOIRAGE

MoIRAGE was developed through a sequence of simpler models and failed formulations. We summarize this progression because several parts of the final model were introduced in response to specific failure modes observed during development. All model-design decisions were made using development data; the held-out UNMC TEST partitions were not used to choose the final architecture.

Table 2: **Evolution of MoIRAGE during development.** Each stage addresses a limitation observed in the previous design.

Stage	What we tried	What we observed	What changed next
1. Original representation	WavLM embeddings released directly.	Clinical information was useful, but participant identity remained strongly recoverable and became easier to link across repeated recordings.	We first tested simple ways to reduce identity information.
2. Simple privacy baselines	Gaussian perturbation and content-only representations.	Noise reduced linkage but also reduced clinical utility. Content-only representations preserved more utility but remained highly linkable.	We moved from perturbation and suppression toward explicit separation of clinical and identity-related information.
3. Adversarial identity suppression	Speaker-fooling and conditional-speaker objectives were used to remove speaker information from the content representation.	These objectives pushed the representation toward collapse and did not give a stable privacy-utility solution.	We stopped adversarially pushing identity out of the content branch and instead gave the identity branch a direct speaker-supervision signal.
4. Factorized representation	The encoder separated content c and identity-related information s, with detached reconstruction and cross-covariance regularization.	Factorization alone was not sufficient; identity information could still remain in the released representation.	We discarded the source identity component and introduced synthetic identity replacement.
5. Synthetic identity replacement	The source component s_{src} was replaced with a sample $\tilde{s}$ from a learned prior before release.	Single-release privacy improved, but identity information could still accumulate when several recordings were combined.	We introduced multi-observation adversarial training.
6. Multi-observation training	A set attacker was trained over several K values. Early privacy objectives used binary classification losses.	Binary losses did not match the ordering behavior of the linkage scores well.	We replaced them with rank, score-gap, and centering objectives over $K \in \{1, 2, 5, 10, 20\}$.
7. Final MoIRAGE	Synthetic identity replacement, repeated-observation training, clinical consistency, stronger anti-collapse regularization, and a utility constraint.	The final model reduced repeated-observation linkage while retaining clinically useful information, although substantial linkage still remained at large K.	Final model used for the held-out evaluation.

C Training Objectives

Two independent released embeddings, $z^{(1)}$ and $z^{(2)}$, are generated from the same rows during each training step. Clinical losses are computed at the participant level. Participant pooling concatenates the mean and standard deviation of that participant's rows.

Clinical prediction loss. Let g_c be the clinical head on pooled content and g_r the clinical head on pooled releases. The supervised clinical loss is

$$\begin{aligned}\mathcal{L}_{\mathrm{clin}} = \; & \mathrm{BCE}(g_c(\bar{c}), y) \\ & + \frac{1}{2}\left[\mathrm{BCE}(g_r(\bar{z}^{(1)}), y) + \mathrm{BCE}(g_r(\bar{z}^{(2)}), y)\right].\end{aligned} \tag{12}$$

Its weight is 0.75.

Clinical teacher loss. We also use out-of-fold probabilities from a balanced logistic-regression teacher with $C = 1.0$ and five inner folds. The squared error between the model probabilities and

the fixed teacher probabilities is applied to the content head and both release heads. This term has weight 2.0.

Clinical consistency. The two synthetic releases should give similar clinical predictions. We use

$$\mathcal{L}_{\text{cons}} = \left\| \sigma(g_r(\bar{z}^{(1)})) - \sigma(g_r(\bar{z}^{(2)})) \right\|_2^2, \tag{13}$$

with weight 2.0. Gradients flow through both releases.

Content bottleneck and prior fitting. The content bottleneck is

$$\mathcal{L}_{\text{VIB}} = \text{KL}\left(q(c \mid x) \,\|\, \mathcal{N}(0, I)\right), \tag{14}$$

with weight 2.5×10^{-4}. The identity-related posterior is matched to the learned prior using

$$\mathcal{L}_{\text{prior}} = \text{KL}\left(q(s \mid x) \,\|\, p(s \mid y)\right), \tag{15}$$

with weight 0.02. Both KL terms are summed over the 32 latent dimensions and then averaged over rows.

Reconstruction. The reconstruction loss combines cosine and mean-squared error:

$$\mathcal{L}_{\text{rec}} = 1 - \cos\left(\hat{x}, \frac{x}{\|x\|_2}\right) + 0.25 \left\| \hat{x} - \frac{x}{\|x\|_2} \right\|_2^2. \tag{16}$$

This term has weight 0.6. The content representation is detached in the reconstruction path.

Content-identity separation. The identity-related branch is trained to match ECAPA speaker embeddings using mean cosine distance, with weight 1.5. We also reduce redundancy between μ_c and μ_s. Each latent dimension is standardized and we compute the 32×32 cross-covariance matrix. The mean squared matrix entry is used as a correlation penalty with weight 0.2.

Multi-observation confusion. For each $K \in \{1, 2, 5, 10, 20\}$, MoIRAGE is trained through a frozen linkage attacker. Each generator update uses 64 set pairs at every K. Let u^+ and u^- denote positive and negative linkage logits and let $\tau = 0.5$. We use

$$\mathcal{L}_{\text{rank}} = \text{Huber}_{0.5}\left(\frac{u_i^+ - u_j^-}{\tau}, 0\right), \tag{17}$$

$$\mathcal{L}_{\text{gap}} = \left(\frac{\bar{u}^+ - \bar{u}^-}{\tau}\right)^2, \qquad \mathcal{L}_{\text{centre}} = \bar{u}^2. \tag{18}$$

The weights are 1.5, 0.75, and 0.02, respectively. The three losses are averaged across the five values of K and multiplied by the privacy ramp.

Anti-collapse regularization. Reducing linkage is not useful if all released embeddings collapse to nearly the same point. We therefore regularize the class-conditional covariance of z. The target variance is $1/d = 1/64$. The loss contains four parts:

$$\begin{aligned} \mathcal{L}_{\text{div}} = {} & 1.25\,\overline{\log^2(\Sigma_{ii}/\text{target})} + 0.2\,\overline{(\Sigma_{ij}/\text{target})^2} \\ & + 2.0\,\overline{\text{ReLU}(0.12\,\text{target} - \lambda_i)/\text{target}} + 2.0\,\|\mathbb{E}[z \mid y]\|_2^2, \end{aligned} \tag{19}$$

plus an effective-rank penalty with weight 4.0. The effective-rank term is activated when the entropy-based rank fraction falls below 0.3. The log-ratio variance penalty is unbounded as variance approaches zero, which makes collapse increasingly expensive.

Utility constraint. After the first 25 epochs, the mean clinical loss over epochs 21-25 is fixed and increased by 5% to define a utility target. Any clinical loss above this target receives an additional penalty:

$$\mathcal{L}_{\text{dual}} = \alpha\ \text{ReLU}(\mathcal{L}_{\text{clin}} - \mathcal{L}_{\text{target}}). \tag{20}$$

The dual variable starts at 1.0, increases with step size 0.5 in proportion to the violation, and is capped at 25.0.

Training schedule. The first 25 epochs focus on utility. The multi-observation privacy terms are then increased linearly from 0 to 1 over the next 60 epochs and remain fully active until epoch 180. All other objectives are active throughout training.

Complete implemented objective. The implementation uses the weighted sum described above:

$$\begin{aligned}\mathcal{L}_{\text{total}} = {} & 0.75\mathcal{L}_{\text{clin}} + 2.0\mathcal{L}_{\text{teacher}} + 2.0\mathcal{L}_{\text{cons}} + 2.5\times10^{-4}\mathcal{L}_{\text{VIB}} \\ & + 0.02\mathcal{L}_{\text{prior}} + 0.6\mathcal{L}_{\text{rec}} + 1.5\mathcal{L}_{\text{spk}} + 0.2\mathcal{L}_{\text{orth}} \\ & + r(t)\,(1.5\mathcal{L}_{\text{rank}} + 0.75\mathcal{L}_{\text{gap}} + 0.02\mathcal{L}_{\text{centre}}) \\ & + \mathcal{L}_{\text{div}} + \mathcal{L}_{\text{dual}},\end{aligned} \tag{21}$$

where $r(t)$ is the privacy ramp. In the teacher-free variant, $\mathcal{L}_{\text{spk}}$ is removed.

D Implementation and Training Details

Optimization. We use two AdamW optimizers with default PyTorch β and ϵ. One optimizer updates the generator, learned prior, and clinical heads. The second updates the linkage attacker. Both use learning rate 10^{-3} and weight decay 10^{-4}. Gradients are clipped to global norm 5.0. We do not use a learning-rate schedule or early stopping. Every model is trained for 180 epochs.

Training budget. One epoch is one full-batch generator update over the training rows of a fold. For fold 0 this contains 2,360 rows. The attacker takes three updates for each generator update. Each attacker update uses 96 balanced set pairs: 48 positive and 48 negative. The generator-side confusion loss additionally uses 64 set pairs for every $K \in \{1, 2, 5, 10, 20\}$.

Checkpoint selection. Checkpoints are evaluated every five epochs using development data only. A checkpoint must satisfy clinical-utility, covariance-trace, effective-rank, and attack thresholds before it is considered feasible. Among feasible checkpoints, we select the checkpoint with the lowest worst-case validation $\text{AUC}^{\star}$. If no checkpoint is feasible, we use the structurally valid checkpoint with the lowest worst-case attack score. TEST data are not used for checkpoint selection.

The fresh-set checks are run every 20 epochs. We also evaluate $K = 7$ during checkpoint selection even though $K = 7$ is never used for adversarial training. This gives one unseen set size during development. The fold-0 final checkpoint is epoch 160.

Random seeds. Training uses seed $42 +$ fold for model initialization, latent sampling, set sampling, and class sampling. Evaluation uses $20260917 +$ fold with separate offsets for different evaluation tasks.

Determinism. All random number generators are seeded, but PyTorch deterministic algorithms and cuDNN deterministic flags are not forced. Results are therefore reproducible within the same environment, but they are not guaranteed to be bitwise identical across platforms.

Hardware. Training and evaluation were run on CPU under Linux 5.15 with Python 3.10.12. The released model is small enough that GPU training is not required.

Training pseudocode.

```
for epoch in 1..180:
    # attacker updates on detached releases
    for step in 1..3:
        K = choice([1,2,5,10,20])
        pairs = sample_balanced_set_pairs(K, batch=96)
        update_attacker(BCE(attacker(pairs), same_participant))

    # factorize input
    h = encoder_backbone(normalize(x))
    mu_c, lv_c, mu_s, lv_s = latent_heads(h)
    s = mu_s + exp(0.5 * lv_s) * eps

    # reconstruction
    x_hat = decoder([stopgrad(mu_c), s])

    # two independent synthetic releases
    z1 = normalize(decoder([mu_c, sample_prior()]))
    z2 = normalize(decoder([mu_c, sample_prior()]))

    # generator privacy loss through frozen attacker
    freeze(attacker)
    multi_loss = mean_over_K([1,2,5,10,20])
    unfreeze(attacker)
```

```
total_loss = clinical + teacher + consistency
           + bottleneck + prior_fit + reconstruction
           + speaker + orthogonality
           + privacy_ramp * multi_loss
           + anti_collapse + utility_dual

update_generator(total_loss)
```

E UNMC COHORT, SPLITS, AND PREPROCESSING

Data provenance. The UNMC cohort is drawn from an IRB-approved longitudinal study of community-dwelling older adults in which naturalistic in-vehicle audio is collected during everyday driving. The source recordings contain spontaneous, multi-speaker conversational speech under real-world conditions, including background noise and variable recording environments. Before feature extraction, we apply automatic speaker diarization and cross-recording speaker linking. The participant speaker is selected using duration-and-coverage criteria, and only segments assigned to that participant are used to construct the analysis chunks. Non-participant speech segments are therefore excluded before WavLM feature extraction. Because diarization and speaker linking are automatic, residual speaker-assignment errors remain possible, and background or vehicle acoustics present during retained participant speech are not explicitly removed. This longitudinal collection provides repeated naturalistic observations from the same participants. All participants provided informed consent for the parent study. The speech recordings are not publicly distributed because they contain sensitive participant data.

Cohort. The processed UNMC cohort contains 137 participants and 3,869 participant-days constructed from 10,286 participant-speech chunks of up to 60 seconds. Of the 137 participants, 127 have the binary clinical label used in this study: 74 are in the cognitive-impairment group and 53 are controls. The remaining 10 participants are retained for privacy evaluation but are excluded from clinical-utility evaluation.

Recording counts are highly unbalanced across participants. The median is 50 chunks per participant, the mean is 75.1, and the range is 2-827. Participant-days have a median of 25, mean of 28.2, and range of 2-64.

Enrollment and probe. Enrollment and probe pools are constructed at the recording level. Recordings from the same participant-day are assigned to the same pool, so enrollment and probe comparisons do not share recordings from the same participant-day.

Outer folds. We use five participant-disjoint outer folds. Their TEST sets are pairwise disjoint and together contain all 137 participants. Each fold contains 86-87 training participants, 23 validation participants, and 27-28 TEST participants. Results in the main paper are pooled across the five outer TEST folds, so each participant is evaluated exactly once by a model that did not train on that participant.

Feature extraction. Participant-assigned speech segments are concatenated into 16-kHz analysis chunks of up to 60 seconds. WavLM-base features are constructed from the last four transformer layers as described in Appendix A.

Train-only preprocessing. The standardizer and 64-component PCA are fitted separately for each fold using TRAIN participants only. The fitted preprocessing objects are then applied unchanged to validation and TEST data. No validation or TEST participant is used to fit the projection.

Missing labels. Rows without a diagnosis label are excluded from clinical-utility calculations but remain available for privacy evaluation.

F BRIDGE2AI DEVELOPMENT REPLICATION

Cohort. The Bridge2AI development cohort used in this study contains 625 participants and 24,144 task-level recordings across 426 task types. Of these, 153 participants have the binary clinical label used here (118 controls and 35 positive cases).

We use five independently generated development splits. Each split contains 500 training and 125 validation participants. These validation sets are not mutually exclusive across folds, although training and validation participants are disjoint within each fold. Across the five validation sets, 155 labeled prediction rows correspond to 113 unique participants, including 29 positive cases. When participant-level clinical performance is reported across folds, repeated predictions for the same participant are averaged before computing the metric.

A separate locked partition of 157 participants was not available for the reported experiments.

Why this is a development replication. The held-out Bridge2AI partition was not available for this experiment. All reported Bridge2AI numbers therefore come from development folds. We use these results only to test whether the same privacy-utility pattern appears in a second cohort. We do not treat them as a second confirmatory test.

Cross-task protocol. UNMC uses recordings from different days. Bridge2AI instead uses *rainbow-passage* recordings for enrollment and *free-speech-1* recordings for probes. The two cohorts therefore test different forms of repeated observation.

Teacher-free variant. The same 64-dimensional MoIRAGE architecture is used, but the ECAPA speaker head and speaker loss are removed because the Bridge2AI representation package used here does not provide the audio needed to build ECAPA targets. No architecture change is made for this cohort.

Feature space. The Bridge2AI vectors used in these experiments are task-level acoustic feature vectors rather than WavLM embeddings. The absolute utility and linkage numbers should therefore not be directly compared with UNMC. The replication is intended to test the behavior of the mechanism across a different cohort and representation space.

G Threat Model and Trial Construction

Attacker knowledge. The attacker has access to released embeddings from evaluation participants, knows the observation count K, and can train attack models using released embeddings from separate training participants whose identities are known. We do not rely on the MoIRAGE algorithm being secret. In the reported standard setting, the attacker does not receive the target participant's raw audio, latent identity-related code, or learned-prior parameters.

Two stronger diagnostics receive additional information. The cross-space attacker is trained using aligned original and released representations from training participants. A separate same-recording diagnostic receives two independent releases of the same recording.

Participant-disjoint attacker training. For each outer fold, attacker training uses only the fold's training participants. These participants are further divided 80/20 by participant for attacker training and early stopping. The attacker is evaluated only on outer TEST participants. No TEST participant is used to train or select an attacker.

Training at each K. Learned attackers are trained at the same K used for evaluation, up to the maximum supported set size. Larger reported K values use all available recordings when a participant does not have enough observations.

Effective set size. For every requested observation count K, we also record the actual number of observations available to the attacker. A result is labeled by a fixed K only when that collection size is supported by the evaluated participants. When availability causes saturation, we report the operating point as *all available* and provide the mean effective set size. This prevents a nominal K from being interpreted as a collection size that was not actually observed.

Positive and negative set trials. A set trial gives the attacker two collections and asks whether they come from the same participant. Batches are balanced between positive and negative trials. For a positive trial, the participant's recordings are randomly permuted and split into two disjoint sets of equal size. If fewer than $2K$ recordings are available, both sets use the largest equal size that can be formed without reusing a row. The same size rule is applied to negative trials. Negative pairs are sampled within the same diagnosis group so that the clinical label alone cannot solve the attack.

Identification trials. Cosine-based identification uses a full enrollment-by-probe score matrix. The diagonal corresponds to the correct participant and off-diagonal entries are incorrect matches. Enrollment and probe come from different recording pools. Set summaries concatenate the mean and standard deviation and are then ℓ_2-normalized.

Trial counts. Across the five UNMC TEST folds, the set and cosine families contain 3,755 trials at each eligible K. The cross-space evaluation contains 5,760 balanced trials.

Sampler checks. During development, we found two problems in an earlier set sampler: positive views could share recordings, and positive and negative trials could have different set sizes. Both could create shortcuts for an attacker. We corrected the sampler, added unit tests for the two conditions, and discarded results produced before the correction. All results in the paper use the corrected sampler.

Release setting attacked. All final TEST attacks use the label-blind MoIRAGE release. Development results that used true-label conditioning are kept separate and are not mixed with the reported TEST numbers.

H ATTACKER ARCHITECTURES

All learned attackers use binary cross-entropy, AdamW with learning rate 10^{-3} and weight decay 10^{-4}, batch size 96 set pairs, and gradient clipping at norm 5. The set encoders are permutation-invariant. For learned pairwise linkage, two set representations a and b are combined using the symmetric feature

$$[\,|a-b|,\ a\odot b\,]. \tag{22}$$

This pair representation is passed to the final linkage head.

Cosine template. This attacker has no learned parameters. The set summary is the ℓ_2-normalized concatenation of the mean and standard deviation. Linkage is the cosine similarity between two summaries.

DeepSets linker. The element encoder ϕ maps $64 \to 64 \to 64$ with GELU. Masked mean and standard-deviation pooling are concatenated, and ρ maps the pooled vector through a 64-dimensional hidden representation with GELU and LayerNorm. This is the attacker family used during MoIRAGE training.

Attention pooling. A single learned query applies additive attention over the set members. The attention-pooled representation is concatenated with the masked mean before the pairwise linkage head.

Statistical MLP. This attacker uses the masked mean, standard deviation, and maximum of the raw set members. These statistics are concatenated and passed through an MLP. It is never used to train MoIRAGE. It becomes the strongest observed attacker for large K.

Cross-space attacker. This attacker uses two separate branches for original and released representations. Each branch maps a 128-dimensional set summary to a 128-dimensional hidden vector with GELU and LayerNorm. The final head receives $[a, b, a \odot b]$. This attacker models a stronger setting in which aligned unprotected and released data are available for training participants.

Capacity selection. Attacker capacity is selected on development folds only. The sweep covers width $\{64, 256\}$, depth $\{2, 3\}$, and training budgets $\{400, 800, 1600\}$ steps for the main learned families at $K = 20$. Increasing capacity gives little validation improvement. The final statistical MLP uses width 64, depth 3, and 1,600 steps. The selected attack settings are frozen before outer TEST evaluation.

Worst-case privacy score. At each K, the reported worst-case linkage score is the maximum across the standard release-linkage attacker families evaluated under the same trial protocol. Diagnostics that use a different trial prior or a different task, such as same-recording and cross-space attacks, are reported separately rather than mixed into this maximum.

I METRICS AND STATISTICAL ANALYSIS

Orientation-invariant AUC. Our primary privacy metric is

$$\mathrm{AUC}^{\star} = \max(\mathrm{AUC}, 1 - \mathrm{AUC}). \tag{23}$$

A value of 0.5 represents chance-level linkage. Values closer to 1 mean that participant identity is easier to recover.

Additional privacy metrics. We also report equal error rate (EER), TPR at 1% and 5% FPR, rank-1 and rank-5 identification, and mean reciprocal rank where available. These metrics show whether the privacy reduction also holds at strict operating points.

Clinical utility. The primary clinical metric is AUROC from logistic regression. We additionally report AUPRC, balanced accuracy, Elastic Net, and linear SVM results.

Bootstrap intervals. For UNMC, clinical-utility intervals use 1,000 participant-level bootstrap resamples over the pooled predictions from the five disjoint outer TEST folds. The paired utility difference uses 2,000 participant-level bootstrap resamples, with the same resampled participant indices applied to both representations. For Bridge2AI, repeated validation appearances of the same participant are first averaged to obtain one prediction per unique participant; confidence intervals are then computed by participant-level bootstrap resampling over the 113 unique evaluated participants.

For learned privacy attackers, confidence intervals use 500 resamples of probe participants within each fold while keeping the enrollment gallery fixed. For example, at $K = 20$, the selected statistical MLP gives $0.811\ [0.779, 0.838]$ for MoIRAGE and $0.963\ [0.948, 0.975]$ for the original embeddings. Because the reported worst-case privacy value is selected as a maximum across attacker families, we do not attach the confidence interval of one selected family to the maximum as if it were an interval for the maximum itself.

Randomness. Release generation and attacker training are both stochastic. Across all five held-out TEST folds, we evaluated the 3×3 combinations of three generator-release seeds and three attacker seeds at $K = 20$, giving 45 fold-seed runs in total. Mean attacker $\mathrm{AUC}^{\star}$ was 0.840, with a pooled between-run standard deviation of 0.039. Resampling the synthetic identity-related component changes the participant's logistic-regression diagnosis probability by 0.024 on average.

J Full Clinical Utility Results

Table 3 reports all three linear clinical classifiers on the 127 labelled UNMC TEST participants. The label-blind release is the primary setting.

Table 3: Clinical AUROC on pooled UNMC TEST.

	Original	Gaussian	Content-only	−replace	−multi	−ECAPA	MoIRAGE
Logistic regression	0.567	0.473	0.615	0.585	0.654	0.634	**0.643**
Elastic Net	0.534	0.489	0.599	0.580	0.680	0.627	0.641
Linear SVM	0.463	0.501	0.637	0.579	0.668	0.644	0.652

For MoIRAGE, logistic-regression AUROC is 0.643 with 95% participant bootstrap interval $[0.549, 0.738]$. AUPRC is 0.706 and balanced accuracy is 0.617. The original embeddings give AUROC 0.567 with interval $[0.465, 0.672]$. The paired AUROC difference is $+0.075$ with interval $[-0.010, 0.163]$. Because this interval includes zero, we describe clinical information as preserved rather than improved.

Prediction stability also depends on the downstream classifier. Across repeated synthetic draws, the average standard deviation of predicted probability is 0.024 for logistic regression, 0.014 for Elastic Net, and 0.090 for the linear SVM.

K Full Baseline and Ablation Results

Table 4: UNMC TEST ablations. Lower linkage is better. Utility is label-blind logistic-regression AUROC.

Variant	$K = 1$	$K = 5$	$K = 20$	$K = 50$	TPR@1%	Rank-1	Utility
Original	0.778	0.930	0.963	0.965	0.474	0.716	0.567
Content-only	0.759	0.913	0.965	0.964	-	-	0.615
MoIRAGE − replacement	0.741	0.887	0.956	0.955	-	-	0.585
MoIRAGE − multi-observation	0.569	0.717	0.875	0.896	-	-	0.654
MoIRAGE − ECAPA	0.530	0.586	0.838	0.861	-	-	0.634
Gaussian $\sigma = 1$	0.528	0.581	0.838	0.860	-	-	0.473
INLP	0.651	0.818	0.908	-	-	-	0.528
MoIRAGE	**0.528**	**0.609**	**0.811**	**0.854**	**0.036**	**0.161**	**0.643**

Gaussian-noise sweep. To characterize the privacy-utility trade-off beyond a single perturbation level, we evaluate Gaussian noise at $\sigma \in \{0.25, 0.5, 1, 2\}$ using the same held-out $K = 20$ linkage protocol and clinical-utility evaluation used for the other baselines. These operating points are shown in Figure 3(b). The $\sigma = 1$ result is also reported in Table 4.

The replacement ablation is the clearest mechanism-level result. Without synthetic replacement, $K = 20$ linkage increases from 0.811 to 0.956, close to the original value of 0.963. The content-only representation also remains highly linkable at 0.965. This shows that separating the representation is not enough by itself.

Removing multi-observation training increases $K = 20$ linkage to 0.875. Removing ECAPA supervision has a smaller effect, increasing linkage to 0.838. At $\sigma = 1$, Gaussian noise reaches the same

Table 5: Attacker-by-attacker AUC* on pooled UNMC TEST.

Representation	Attacker	$K=1$	$K=2$	$K=5$	$K=10$	$K=20$	$K=30$	$K=50$
Original	cosine template	0.772	-	0.926	-	0.935	-	0.934
Original	DeepSets	0.767	-	0.897	-	0.939	-	0.946
Original	statistical MLP	0.778	-	0.930	-	0.963	-	0.965
Original	cross-space	0.760	-	0.928	-	0.935	-	-
MoIRAGE	cosine template	0.528	0.528	0.581	0.624	0.620	0.620	0.620
MoIRAGE	DeepSets	0.505	0.548	0.588	0.649	0.659	0.674	0.658
MoIRAGE	attention pooling	0.503	-	0.609	-	0.666	-	0.675
MoIRAGE	statistical MLP	0.528	0.559	0.580	0.730	0.811	0.848	0.854
MoIRAGE	cross-space	0.515	-	0.525	-	0.544	-	-

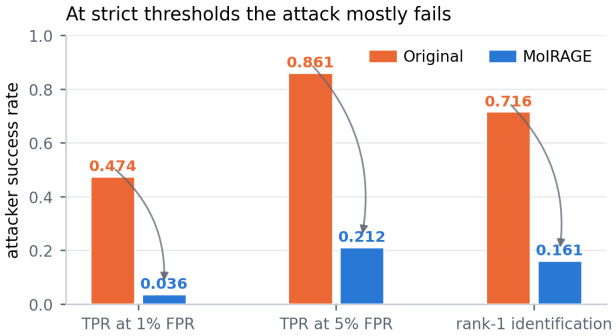


Figure 6: **Participant linkage at strict operating points.** At $K = 20$, MoIRAGE substantially reduces attacker success at 1% FPR, 5% FPR, and rank-1 participant identification compared with the original embeddings.

$K = 20$ linkage as the no-ECAPA variant but reduces clinical AUROC to 0.473. This shows that low linkage can also be obtained by damaging the representation, which is not the goal of MoIRAGE.

INLP baseline. We implement Iterative Nullspace Projection (INLP) (Ravfogel et al., 2020) as an established linear concept-erasure baseline. For each outer fold, INLP is fitted using TRAIN participants only on the 64-dimensional WavLM-PCA representation. Starting from the identity projection, we repeatedly fit a multinomial logistic-regression classifier to predict participant identity and remove the leading right-singular direction of its weight matrix. We use rank-one removal at each iteration because the number of participant classes exceeds the representation dimension.

The number of removed directions is selected using validation data only. We evaluate 16, 32, and 48 removed directions; 32 directions satisfy the predefined utility-feasibility criterion and give the lowest validation $K = 20$ attack score among the feasible settings. The selected projection is then applied unchanged to the corresponding outer TEST partition.

On pooled outer TEST, INLP gives linkage AUC* values of 0.651, 0.818, and 0.908 at $K = 1$, 5, and 20, respectively, with clinical AUROC 0.528. At the all-available operating point, linkage is 0.916.

L Repeated-Observation Results in Full

The attacker used during training is not the strongest attacker at evaluation. At $K = 20$, DeepSets reaches 0.659 while the statistical MLP reaches 0.811. This is why we report the strongest observed attacker instead of reporting only the training adversary.

The attacker gap also grows with K. At $K = 1$, all standard MoIRAGE attackers are close to chance. At larger K, the statistical MLP can combine weak identity information across recordings more effectively.

At $K = 20$, the strongest attacker changes from EER 0.084 to 0.271, TPR@5% FPR from 0.861 to 0.212, TPR@1% FPR from 0.474 to 0.036, rank-1 from 0.716 to 0.161, rank-5 from 0.956 to 0.577, and mean reciprocal rank from 0.824 to 0.349 when moving from the original embeddings to MoIRAGE.

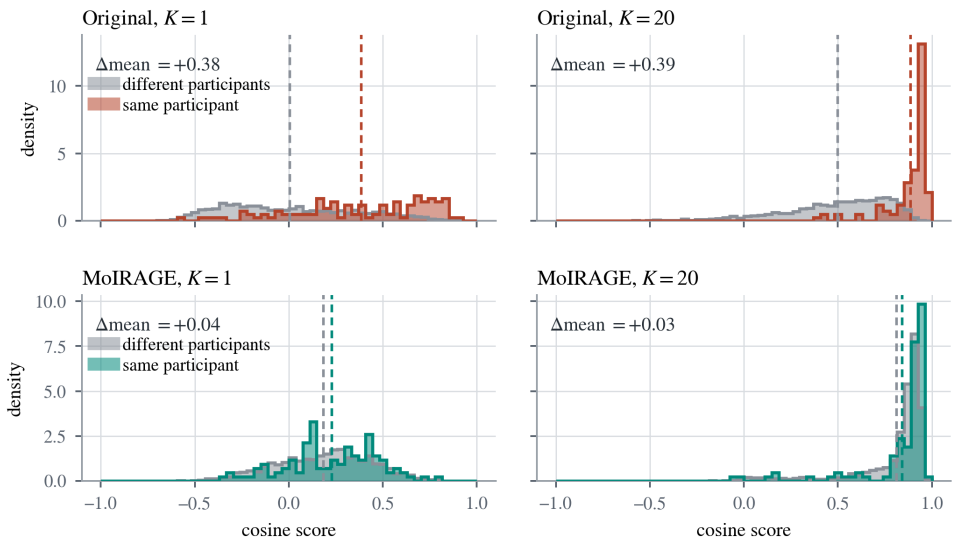


Figure 7: **Cosine-linkage score distributions before and after MoIRAGE.** Same-participant and different-participant comparisons are shown for the original and MoIRAGE representations at $K = 1$ and $K = 20$. Dashed lines indicate distribution means. MoIRAGE substantially reduces the separation between same- and different-participant scores. This cosine-based diagnostic complements the stronger learned attackers used for the primary privacy evaluation.

M Why Repeated Observations Recover Residual Identity

We use a simplified random-effects analysis to understand why participant linkage increases when several released recordings are combined. This analysis is intended to explain the empirical behavior of MoIRAGE; it is not a formal privacy guarantee.

Let

$$z_{ijr} = \mu_i + \delta_{ij} + \epsilon_{ijr},$$

where μ_i represents participant-stable information, δ_{ij} represents variation between recordings from participant i, and ϵ_{ijr} represents stochastic variation caused by independently sampling the synthetic identity-related component for release r.

For repeated synthetic releases of the same recording, i and j remain fixed. If ϵ_{ijr} has covariance Σ_{rel} and is independent across releases, then

$$\text{Cov}\left(\frac{1}{R}\sum_{r=1}^{R}\epsilon_{ijr}\right) = \frac{\Sigma_{\text{rel}}}{R}.$$

Thus, stochastic variation introduced by synthetic replacement becomes easier to average out as more releases are observed.

We test this prediction directly by repeatedly releasing the same recording with independent synthetic identity samples. For each recording and release count R, we average R independently generated releases and quantify dispersion as the trace of the empirical covariance of these averaged representations. Across the five held-out folds, mean dispersion decreases from 0.711 for one release to 0.355 for two, 0.140 for five, 0.069 for ten, and 0.034 for twenty releases. A power-law fit gives approximately $R^{-1.03}$, closely matching the expected $1/R$ rate. The $1/R$ prediction gives $R^2 > 0.9999$ in each fold.

A variance decomposition provides additional context. Averaged across folds, replacement-draw variation accounts for approximately 82.9% of the observed variance, recording-within-participant variation for 14.8%, and stable participant variation for 2.3%. Although the participant component is small for an individual release, it does not disappear when independent stochastic variation is averaged across observations.

This interpretation is also consistent with the ablation results. The content-only representation remains highly linkable at $K = 20$ ($\text{AUC}^{\star} = 0.965$), showing that participant information remains outside the replaced identity component. In the cosine-similarity diagnostic, the mean same-participant versus different-participant margin changes from -0.025 at $K = 1$ to $+0.051$ at $K = 20$,

Table 6: **Variance decomposition of MoIRAGE releases.** Values are mean shares across the five held-out folds.

Component	Variance share
Synthetic replacement draw	82.9%
Recording within participant	14.8%
Participant-stable component	2.3%

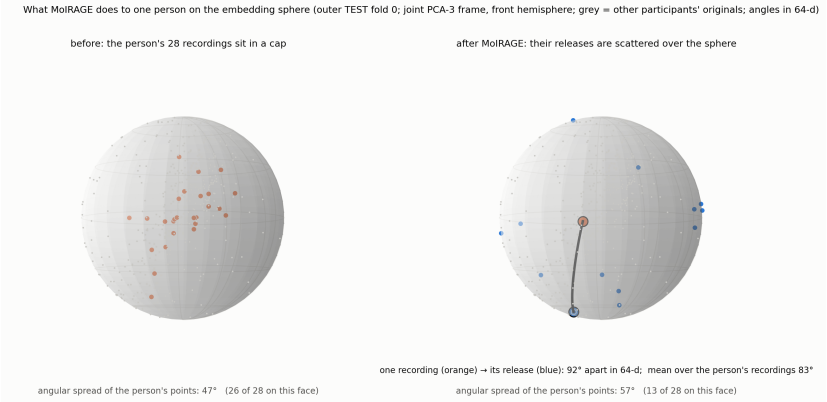


Figure 8: **Participant-level embedding geometry before and after MoIRAGE.** The visualization illustrates how one participant's recordings change after release. It is included as an illustrative analysis rather than as evidence of formal privacy.

and is positive in four of five folds at $K = 20$. We therefore interpret the remaining large-K leakage as consistent with stable participant information becoming more visible as stochastic release variation is averaged away.

N Additional Visualizations

The main paper focuses on the repeated-observation curve, replacement ablation, and strict operating points. Additional geometry visualizations are useful for checking that lower linkage is not caused by complete representation collapse.

On UNMC TEST, the class-conditional covariance trace of MoIRAGE releases is 0.81, compared with 1.0 for an isotropic distribution. Effective rank is 23 of 64 dimensions, compared with 19 for the original embeddings. These results do not support complete representation collapse. Gaussian noise at $\sigma = 1$ has effective rank 62 but much lower clinical utility, showing that simply spreading the representation is not enough.

O Bridge2AI Results in Full

At $K = 20$, TPR at 1% FPR falls from 0.920 to 0.030 and EER changes from 0.027 to 0.382. The teacher-free variant also retains clinical information on the development folds. These results support the same overall pattern seen on UNMC, but they are development results and should not be interpreted as an independent confirmatory test.

Clinical utility across conditions. We additionally examine whether clinical information is preserved consistently across Bridge2AI conditions. The analysis uses the same train-to-held-out-validation protocol as the primary Bridge2AI evaluation. Across the 12 conditions with sufficient validation support, preservation is strongly task-dependent. Cognitive impairment increases from 0.924 to 0.970 mean AUROC, while Parkinson's disease increases from 0.867 to 0.899. The remaining conditions decrease after transformation. The largest reductions occur for muscle tension dysphonia (0.865 to 0.575), airway stenosis (0.904 to 0.713), and PTSD (0.774 to 0.606). These results indicate that MoIRAGE does not preserve arbitrary clinical information equally; the largest losses are concentrated among conditions whose diagnostic signal may depend strongly on voice-quality characteristics.

Conditions represented in only one development fold are omitted from this summary table and retained in the released evaluation outputs.

Table 7: Bridge2AI development replication with the teacher-free MoIRAGE variant. Lower $\mathrm{AUC}^{\star}$ is better.

Representation	Attacker	$K = 1$	$K = 5$	$K = 20$	$K = 50$
Original	cosine template	0.928	0.935	0.935	0.935
Original	DeepSets	0.799	0.970	0.996	0.998
Original	statistical MLP	0.818	0.969	0.996	0.997
Original	cross-space	0.929	0.935	0.935	-
MoIRAGE	cosine template	0.557	0.575	0.575	0.575
MoIRAGE	DeepSets	0.510	0.610	0.709	0.724
MoIRAGE	statistical MLP	0.508	0.601	0.677	0.778
MoIRAGE	cross-space	0.536	0.554	0.554	-

Table 8: **Bridge2AI participant-level clinical utility.** AUROC is computed over 113 unique labeled participants appearing in at least one held-out validation fold. Because validation folds overlap, prediction scores from repeated appearances of the same participant are averaged before metric computation. Confidence intervals use participant-level bootstrap resampling.

Representation	AUROC	95% CI
Original	0.913	[0.820, 0.974]
MoIRAGE (label-blind, multi-draw)	**0.948**	[0.862, 0.996]

Table 9: **Bridge2AI clinical utility across auxiliary conditions.** For each condition, models are fitted on the outer-fold training participants and evaluated on the corresponding held-out validation participants, matching the primary Bridge2AI evaluation protocol. Values are mean AUROC across eligible folds. The primary cognitive-impairment result in Section 6 instead aggregates repeated validation predictions to one score per unique participant before computing pooled participant-level AUROC, so the numerical values are not expected to be identical. Condition-specific controls are obtained from the corresponding phenotype tables.

Condition	Original	MoIRAGE	Folds
Cognitive impairment	0.924	**0.970**	5
Parkinson's disease	0.867	**0.899**	5
Psychiatric history	0.744	0.640	5
Unilateral vocal-fold paralysis	0.853	0.742	5
Unexplained chronic cough	0.820	0.703	4
Laryngeal dystonia	0.895	0.761	5
Benign lesions	0.758	0.622	5
Depression	0.816	0.675	5
ADHD (adult)	0.775	0.612	5
PTSD (adult)	0.774	0.606	5
Airway stenosis	0.904	0.713	5
Muscle tension dysphonia	0.865	0.575	5

P Failure Cases and Limitations

Identity is reduced, not removed. MoIRAGE does not solve repeated-observation identity leakage. On UNMC TEST, the strongest attacker still reaches $\mathrm{AUC}^{\star} = 0.811$ at $K = 20$ and 0.854 at $K = 50$. Rank-1 identification at $K = 20$ remains 0.161. These values are lower than the original embeddings but remain far from chance.

Risk is not the same for every participant. Some participants remain easier to identify after release than others. The current work does not explain which speaker or recording properties cause this difference. Understanding these participant-level failure cases is an important direction for future work.

Risk increases with the number of recordings. Repeated observations continue to increase linkage even after synthetic identity replacement. The aggregation analysis in Appendix M provides one explanation: independent release variation averages out approximately as $1/R$, while stable participant information does not. This does not establish a formal privacy bound, but it shows why a deployment that releases many recordings from one participant should not use the single-recording result as its privacy estimate.

Clinical utility is preserved on average. The clinical AUROC difference has a confidence interval that includes zero. We therefore do not claim that MoIRAGE improves diagnosis. We only claim that clinically useful information is retained while identity linkage is reduced.

No formal privacy guarantee. MoIRAGE provides empirical resistance to the attacks evaluated here. It does not provide anonymity, differential privacy, or a formal upper bound on re-identification risk.

Scope. The current study uses one main clinical task, two cohorts, and one input feature family per cohort. MoIRAGE operates on embeddings rather than raw audio. Future work should study additional clinical tasks, speech encoders, longer longitudinal histories, stronger adaptive attackers, and formal privacy mechanisms.

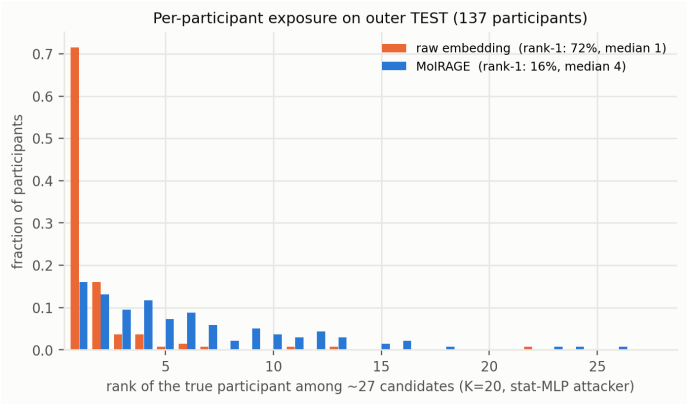


Figure 9: **Participant-level identification risk at $K = 20$.** The distribution of true-participant retrieval ranks shows that MoIRAGE reduces identification success for many participants, while some participants remain easier to identify than others. This heterogeneity motivates participant-level analysis of residual privacy risk.

Q REPRODUCIBILITY

Frozen artifacts. A manifest stores SHA-256 hashes for the training scripts, evaluation scripts, configurations, split files, preprocessing objects, and checkpoints. The architecture, loss functions, identity-replacement mechanism, dimensions, checkpoint-selection rule, and privacy evaluation protocol were frozen before the final TEST evaluation.

Software. Experiments use Python 3.10.12, PyTorch 2.14.0, scikit-learn 1.7.2, NumPy 2.2.6, SciPy 1.15.3, and pandas 2.3.3 on Linux 5.15.

Seeds. Training uses base seed 42, final evaluation uses base seed 20260917. Fold-specific offsets are added as described in Appendix D.

Test-set discipline. The outer UNMC TEST folds are used only after the architecture, model-selection rule, attacker families, and attacker capacities are fixed on development data. Figures and tables are generated from stored evaluation outputs rather than from repeated TEST tuning.
The Gaussian-noise sweep and INLP comparison were added after the MoIRAGE design and evaluation protocol were frozen. INLP direction count was selected using validation data only, and neither baseline affected MoIRAGE model, checkpoint, or attacker selection.

Code and data. Our reproducibility artifacts include training and evaluation scripts, configuration files, split metadata, frozen manifests, per-fold outputs, aggregated result tables, and figure-generation scripts. UNMC recordings cannot be redistributed by the authors. Bridge2AI data remain subject to the access conditions of the dataset maintainers.