SILO: SUBSPACE-PRESERVING SPECTRAL INJECTION FOR GENERALIZABLE AUDIO DEEPFAKE DETECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Audio deepfake detectors often struggle to generalize across unseen synthesis systems and real-world acoustic conditions. We first analyze the **complementary failure modes** of classical spectral features and self-supervised learning (SSL) representations: spectral cues expose transferable synthesis artifacts but are sensitive to acoustic variation, whereas SSL representations are more robust to such variation yet can miss fine-grained spoofing artifacts. Motivated by this observation, we propose **Spectral Injection via Low-rank Orthogonality (SILO)**, which uses utterance-adaptive spectral evidence to complement the robust SSL representation through a dedicated low-rank spectral-injection branch. SILO further introduces a **geometric constraint** that separates low-rank directions from their signed amplitudes and a **preserved-subspace constraint** that protects selected high-energy pretrained directions. Both constraints are maintained through optimization on restricted Stiefel manifolds. Across cross-dataset evaluations, SILO outperforms state-of-the-art methods, achieving 1.02% EER on ASVspoof 2019, 1.56% on ASVspoof 2021, and 14.38% macro-EER on MLAAD. We further introduce **Voice-Disorder Deepfake (VDD)** and **ForVoice Deepfake (FVD)**, two evaluation datasets targeting pathological speech and long-form forensic recordings, respectively. These datasets extend evaluation to these underexplored conditions and reveal substantial remaining challenges in pathological-speech and long-form deepfake detection.

1 INTRODUCTION

Generalization remains a central challenge of audio deepfake detection (ADD): detectors that perform well on synthesis systems observed during training can degrade substantially when confronted with unseen generators, languages, or recording conditions (Müller et al., 2022; Todisco et al., 2019; Yamagishi et al., 2021; Müller et al., 2024). These shifts are often grouped as out-of-distribution generalization, but they impose different requirements on a detector. Cross-generator transfer relies on artifacts shared across synthesis methods, whereas robustness to real-world recordings requires invariance to background noise, room reverberation, microphones, and transmission channels. Treating both settings as a single out-of-distribution problem can therefore obscure an important tension between *cross-generator generalization* and *robustness to uncontrolled acoustic conditions*.

Our analysis of existing ADD models reveals complementary failure modes between classical spectral and self-supervised learning (SSL) detectors, as illustrated in Fig. 1. Spectral detectors capture fine spectral and harmonic artifacts that often transfer to unseen synthesis systems, including recent LLM-based generators such as Qwen3-TTS and MOSS-TTS, but are sensitive to uncontrolled acoustic variation. In contrast, SSL-based detectors are more robust to such variation, yet may overlook generator artifacts that transfer across synthesis families. This observation suggests our first design prin-

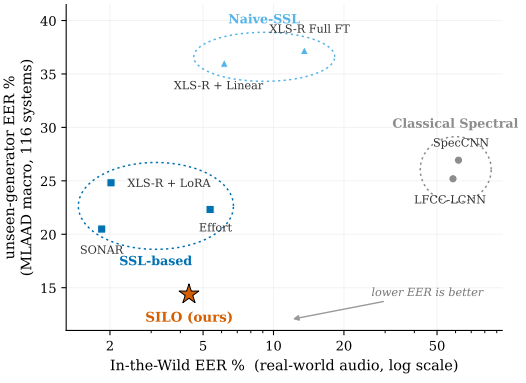


Figure 1: **Complementary generalization behavior of ADD detector families and SILO.**

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.

ciple: *spectral evidence should complement the robust SSL representation rather than replace it*. Importantly, the amount of spectral evidence used by the detector should depend on the input, since cues that are informative in clean speech may become unreliable under noise or reverberation.

A second challenge concerns how the SSL model should be adapted. Recent state-of-the-art (SOTA) ADD methods (Wang et al., 2023; Wu et al., 2024; Laakkonen et al., 2026; 2025) increasingly use parameter-efficient fine-tuning (PEFT), particularly Low-Rank Adaptation (LoRA) Hu et al. (2022). For a frozen projection W_0, standard LoRA learns a low-rank residual $\Delta W = BA$ to capture *forgery-relevant* features. While parameter-efficient, the unconstrained factors B and A provide no explicit control over their strength or the pretrained directions they can modify. This becomes particularly important when incorporating external spectral evidence: stronger adaptation may help the model detect synthesis artifacts, but it can also weaken the robust representations learned during pretraining. This leads to our second design principle: *adaptation should explicitly control both how strongly the model is modified and which pretrained directions are allowed to change*.

Motivated by these two principles, we propose **Spectral Injection via Low-rank Orthogonality (SILO)**. SILO retains the frozen SSL backbone and introduces two low-rank residual branches at selected attention projections: an *SSL-adaptation branch* that adapts the pretrained representation for deepfake detection, and a *spectral-injection branch* that introduces complementary spectral evidence. We also introduce an *utterance-level gate* to control the contribution of the spectral branch for each input, allowing the detector to emphasize spectral artifacts when they are useful and rely more strongly on the robust SSL representation when they are not. SILO further constrains both low-rank branches to make the adaptation itself controllable. First, the **geometric constraint** uses a three-factor parameterization inspired by the singular value decomposition (SVD) (Fig. 2), $\Delta W = U\,\mathrm{diag}(\alpha)V^\top$, with orthonormal factors on Stiefel manifolds (Li et al., 2026; Lion et al., 2026). This separates adaptation directions from their signed amplitudes, making $|\alpha_i|$ the exact singular values of the learned residuals. Second, the **preserved-subspace constraint** prevents the learned residuals from modifying selected high-energy directions of the frozen projection. Together, the two constraints determine *how strongly* the model is adapted and *which pretrained directions* can be modified, while the utterance-level gate determines *when* spectral evidence should contribute. We maintain these constraints throughout training by optimizing the corresponding factors directly on restricted Stiefel manifolds.

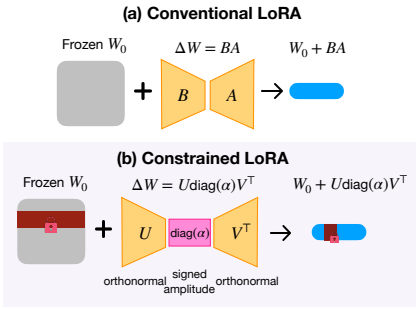


Figure 2: **Overview of SILO's constrained low-rank adaptation.**

Beyond model design, current ADD benchmarks leave important deployment conditions underexplored. Widely used benchmarks such as In-the-Wild and the ASVspoof families primarily contain relatively short recordings of healthy speech (Müller et al., 2022; Yamagishi et al., 2021; Todisco et al., 2019). Therefore, it remains unclear how current detectors behave when genuine speech contains atypical voice characteristics or when spoofing evidence must be identified over substantially longer recordings. To study these settings, we construct two additional evaluation datasets. **Voice-Disorder Deepfake (VDD)** is based on bonafide pathological speech (Jenei et al., 2021) and evaluates whether pathology-related voice characteristics are mistaken for synthesis artifacts. **ForVoice Deepfake (FVD)** derives bonafide recordings from the ForVoice120 forensic-voice corpus (Sztahó & Fejes, 2023) and enables evaluation on long-form recordings.

Contributions. Our main contributions are summarized as follows:

- We introduce **SILO**, a two-branch adaptation method that retains a frozen SSL backbone while injecting complementary spectral evidence through a dedicated low-rank branch.
- We formulate **geometric and preserved-subspace constraints** that explicitly control the strength and direction of adaptation, and optimize them directly on restricted Stiefel manifolds.
- We introduce **VDD and FVD**, two evaluation ADD datasets for evaluating ADD under pathological speech and long-form forensic recordings, respectively.

2 Related work

Generalized Audio Deepfake Detection. Audio deepfake detection (ADD) has progressed from spectral front-ends with lightweight classifiers (Wu et al., 2020) to raw-waveform architectures (Tak et al., 2021; Jung et al., 2022; Chen et al., 2024) and large self-supervised speech encoders (Nguyen-Le et al., 2026; Tak et al., 2022; Wang & Yamagishi, 2024). Despite strong in-domain performance, cross-dataset generalization remains difficult because unseen generators, languages, codecs, and recording conditions can induce substantially different shifts (Müller et al., 2022; 2024). Recent approaches improve generalization through augmentation and domain-generalization objectives (Laakkonen et al., 2026), or adapt pretrained SSL models parameter-efficiently using LoRA and related adapters (Wang et al., 2023; Wu et al., 2024; Laakkonen et al., 2025). SILO instead exploits the complementary behavior of SSL and spectral representations: it retains the robust pretrained SSL pathway while injecting fine-grained spectral evidence through an utterance-adaptive, constrained low-rank branch.

Evaluation Datasets. Cross-dataset ADD evaluation commonly uses WaveFake (Frank & Schönherr, 2021), ASVspoof 2019/2021 (Todisco et al., 2019; Yamagishi et al., 2021), In-the-Wild (Müller et al., 2022), and MLAAD (Müller et al., 2024), covering complementary shifts in generators, recording conditions, codecs, and languages. However, these benchmarks do not explicitly target pathological speech or recording-level detection over multi-minute audio. Therefore, we construct two *evaluation deepfake datasets*: VDD, based on Hungarian pathological-speech resources (Jenei et al., 2021), and FVD, based on the ForVoice120 forensic-voice corpus (Sztahó & Fejes, 2023). Both use a common panel of 7 TTS/VC attacks, enabling controlled evaluation under these previously underexplored conditions. Full dataset comparisons are provided in Appendix E.1.

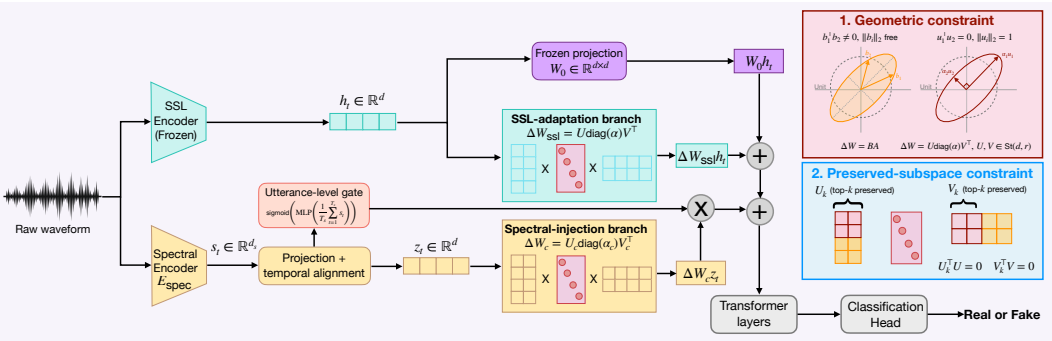


Figure 3: **Overview of SILO**. At each adapted projection, SILO combines the frozen SSL projecttion W_0 with a constrained SSL-adaptation residual and an utterance-gated spectral-injection residual.

3 Spectral Injection via Low-rank Orthogonality (SILO)

3.1 Design Objectives

SILO is designed around two objectives motivated by the complementary behavior of spectral and SSL representations discussed in Sec. 1. **First**, we aim to supplement the robust SSL representation with fine-grained spectral evidence when that evidence is useful for detecting unseen synthesis artifacts. **Second**, we aim to adapt the pretrained SSL model without freely modifying the pretrained directions that may contribute to its robustness. SILO addresses these two objectives through utterance-adaptive spectral injection and constrained low-rank adaptation, respectively.

Let x be an input waveform and $\mathcal{J}$ the index set of adapted self-attention projections. For each $\ell \in \mathcal{J}$, let $W_0^{(\ell)} \in \mathbb{R}^{d\times d}$ be a frozen projection and $h_t^{(\ell)} \in \mathbb{R}^d$ be SSL input at time $t \in \{1, \dots, T\}$, where d is the common hidden width. SILO combines this frozen projection with two low-rank components: an *SSL-adaptation residual*, which acts on the SSL representation, and a *spectral-injection map*, which introduces a time-aligned spectral representation $z_t \in \mathbb{R}^d$. We omit the projection index (ℓ) when describing a single projection. Fig. 3 illustrates the method.

3.2 HIGH-RESOLUTION SPECTRAL BRANCH

A frozen spectral encoder produces $s_{1:T_s} = E_{\text{spec}}(x)$, where $s_j \in \mathbb{R}^{d_s}$, T_s is the spectral sequence length, and d_s is the encoder output width. A learned projection $P_z : \mathbb{R}^{d_s} \to \mathbb{R}^d$ maps the features to the SSL hidden width, followed by token-wise normalization and temporal alignment:

$$(z_t)_{t=1}^{T} = \mathcal{A}_{T_s \to T}\Big((\mathrm{LN}(P_z s_j))_{j=1}^{T_s}\Big), \qquad z_t \in \mathbb{R}^d. \tag{1}$$

Here, P_z changes the feature dimension, while $\mathcal{A}_{T_s \to T}$ changes only temporal resolution. The SILO injection mechanism does not depend on a particular spectral front-end. In our implementation, E_{spec} operates on Linear Frequency Cepstral Coefficients (LFCC) (Sahidullah et al., 2015). We analyze alternative classical spectral front-ends in Appendix F.6.

3.3 UTTERANCE-ADAPTIVE SPECTRAL INJECTION

At each selected projection, SILO combines a low-rank SSL-adaptation residual with an utterance-adaptive spectral injection. Let r and r_c denote the SSL-adaptation and spectral-injection ranks, respectively, with $1 \le r, r_c \ll d$. We parameterize two low-rank components as $\Delta W_{\text{ssl}} := U D_\alpha V^\top$ and $\Delta W_c := U_c D_{\alpha_c} V_c^\top$, where $D_\alpha := \mathrm{diag}(\alpha)$, $D_{\alpha_c} := \mathrm{diag}(\alpha_c)$, $U, V \in \mathbb{R}^{d \times r}$, $U_c, V_c \in \mathbb{R}^{d \times r_c}$, $\alpha \in \mathbb{R}^r$, and $\alpha_c \in \mathbb{R}^{r_c}$. Because spectral evidence can become unreliable under noise or other acoustic variation, SILO controls its contribution at the utterance level. We first pool the spectral sequence and compute $g_\phi(x) = \rho\left(a_\phi\left(\frac{1}{T_s}\sum_{j=1}^{T_s} s_j\right)\right) \in (0,1)$, where $a_\phi : \mathbb{R}^{d_s} \to \mathbb{R}$ is learned and $\rho(u) = (1+e^{-u})^{-1}$ is the sigmoid function. The gate $g_\phi(x)$ determines how strongly spectral evidence contributes to the current utterance. For a gate value β, the resulting SILO projection is:

$$\mathcal{F}_\beta(h_t, z_t) = \underbrace{W_0 h_t}_{\text{frozen SSL}} + \underbrace{\Delta W_{\text{ssl}} h_t}_{\text{SSL residual}} + \underbrace{\beta(x)\Delta W_c z_t}_{\text{spectral injection}} \tag{2}$$

We apply this construction to the selected query, key, value, and output projections.

3.4 GEOMETRIC AND PRESERVED-SUBSPACE CONSTRAINTS

SILO introduces two additional constraints: (1) *geometric constraint* controls the strength of adaptation by placing the low-rank factors of both residual branches on the Stiefel manifold; and (2) *preserved-subspace constraint* requires the learned directions to lie outside selected high-energy subspaces of the frozen projection.

Geometric constraint. For $1 \le s \le d$, let $\mathrm{St}(d,s) := \{X \in \mathbb{R}^{d \times s} : X^\top X = I_s\}$ denote the Stiefel manifold with orthonormal columns. SILO requires:

$$U, V \in \mathrm{St}(d, r), \qquad U_c, V_c \in \mathrm{St}(d, r_c). \tag{C1}$$

Constraint C1 makes the columns of the direction factors orthonormal. For a matrix M, let $|M|_{S_q}$ denote its Schatten-q norm, i.e., the ℓ_q norm of its singular values. For vectors, $|\cdot|_q$ denotes the usual ℓ_q norm.

Proposition 1 (Exact singular values and residual bound)**.** *Under Constraint C1, then:* ***(i)*** *The singular values of ΔW_{ssl} are $\{|\alpha_i|\}_{i=1}^r$ and $d-r$ zeros, and those of ΔW_c are $\{|\alpha_{c,i}|\}_{i=1}^{r_c}$ and $d-r_c$ zeros;* ***(ii)*** *For every $q \in [1,\infty]$, $\|\Delta W_{\text{ssl}}\|_{S_q} = \|\alpha\|_q$ and $\|\Delta W_c\|_{S_q} = \|\alpha_c\|_q$;* ***(iii)*** *For every $\beta \in \mathbb{R}$, the two-input residual operator $\mathcal{R}_\beta : \mathbb{R}^d \oplus \mathbb{R}^d \to \mathbb{R}^d$, defined by $\mathcal{R}_\beta(h,z) := \Delta W_{\text{ssl}} h + \beta \Delta W_c z$, satisfies $\|\mathcal{R}_\beta(h,z)\|_2 \le \|\alpha\|_\infty \|h\|_2 + |\beta|\,\|\alpha_c\|_\infty \|z\|_2$ for all $h, z \in \mathbb{R}^d$, and hence $\|\mathcal{R}_\beta\| \le (\|\alpha\|_\infty^2 + \beta^2 \|\alpha_c\|_\infty^2)^{1/2}$ for the operator norm induced by $\|(h,z)\| := (\|h\|_2^2 + \|z\|_2^2)^{1/2}$.*

Proposition 1 (Appendix C.3) shows that α and α_c directly control the residual strengths, while spectral injection is further modulated by β and $\|z\|_2$.

Preserved-subspaces constraint. To identify the pretrained directions to preserve, we consider a fixed SVD of the frozen projection, $W_0 = \bar{U}\,\mathrm{diag}(\sigma)\bar{V}^\top$ $(\sigma_1 \ge \cdots \ge \sigma_d \ge 0)$. The singular values rank the pretrained directions by their energy. Given a *preserved energy ratio* $\pi \in [0,1)$, we choose the smallest number of leading directions whose squared singular values account for at least a fraction π of the total energy $k := \min\left\{j \in \{0,\dots,d\} : \sum_{i=1}^{j} \sigma_i^2 \ge \pi \sum_{i=1}^{d} \sigma_i^2\right\}$. We then define

$U_k := \bar{U}_{[:,1:k]},\ V_k := \bar{V}_{[:,1:k]}$, whose columns span the protected output and SSL-input subspaces, respectively.

SILO requires the learned low-rank directions to be orthogonal to these protected subspaces:

$$\underbrace{U_k^\top U = 0}_{(A)}, \qquad \underbrace{V_k^\top V = 0}_{(B)}, \qquad \underbrace{U_k^\top U_c = 0}_{(C)}. \tag{C2}$$

The (A) and (C) conditions prevent the SSL-adaptation and spectral-injection residuals from modifying the protected output directions: $U_k^\top \Delta W_{\text{ssl}} = 0,\ U_k^\top \Delta W_c = 0$. Similarly, (B) implies $\Delta W_{\text{ssl}} V_k = 0$, so the SSL-adaptation residual leaves the protected SSL-input directions unchanged. We do not impose a corresponding constraint on V_c, since V_c acts on the external spectral representation z_t rather than the SSL representation h_t.

For any orthonormal $Q \in \mathbb{R}^{d\times k}$ with $Q^\top Q = I_k$, we define the *restricted Stiefel manifold* $\text{St}_Q(d,s) := \{X \in \mathbb{R}^{d\times s} : X^\top X = I_s,\ Q^\top X = 0\}$. Constraints (C1)–(C2) can therefore be written as $U \in \text{St}_{U_k}(d,r)$, $V \in \text{St}_{V_k}(d,r)$, $U_c \in \text{St}_{U_k}(d,r_c)$, and $V_c \in \text{St}(d,r_c)$.

Theorem 1 (Exact preservation)**.** *Under constraint C2, let $W_{\text{adapt}} := W_0 + \Delta W_{\text{ssl}}$ denote the adapted SSL projection. Partition the full singular bases as $\bar{U} = [\,U_k\ U_\perp\,]$ and $\bar{V} = [\,V_k\ V_\perp\,]$, and define $\Sigma_k := \text{diag}(\sigma_1,\dots,\sigma_k),\ \Sigma_\perp := \text{diag}(\sigma_{k+1},\dots,\sigma_d)$. Then:*

(i) For every $h, z \in \mathbb{R}^d$ and every $\beta \in \mathbb{R}$, $U_k^\top \mathcal{F}_\beta(h,z) = \Sigma_k V_k^\top h$.

(ii) For every $v \in \text{range}(V_k)$, $W_{\text{adapt}} v = W_0 v$.

(iii) If $\bar{u}_i$ and $\bar{v}_i$ are the ith columns of $\bar{U}$ and $\bar{V}$, then, for every $i \le k$, $W_{\text{adapt}}\bar{v}_i = \sigma_i \bar{u}_i,\ W_{\text{adapt}}^\top \bar{u}_i = \sigma_i \bar{v}_i$.

(iv) In the pretrained singular-vector bases,

$$\begin{bmatrix} U_k^\top \\ U_\perp^\top \end{bmatrix} W_{\text{adapt}} [V_k\ V_\perp] = \begin{bmatrix} \Sigma_k & 0 \\ 0 & \Sigma_\perp + U_\perp^\top \Delta W_{\text{ssl}} V_\perp \end{bmatrix}. \tag{3}$$

Theorem 1 (proof in Appendix C.4) formalizes the main purpose of Constraint C2: SILO can learn new SSL and spectral directions while leaving the selected pretrained singular directions exactly unchanged.

3.5 Optimization on the restricted Stiefel manifolds

Constraints C1 and C2 must hold throughout training. Rather than encouraging them through soft regularization, SILO optimizes the constrained factors directly on their feasible manifolds (Absil et al., 2008). For one adapted projection, the feasible set is $\mathcal{M} := \text{St}_{U_k}(d,r) \times \text{St}_{V_k}(d,r) \times \text{St}_{U_k}(d,r_c) \times \text{St}(d,r_c) \times \mathbb{R}^r \times \mathbb{R}^{r_c}$. This set is nonempty when $k \le d - \max(r, r_c)$, which ensures that enough dimensions remain outside the protected subspaces for both low-rank branches.

Training objective. Let θ denote all trainable parameters, including all constrained low-rank factors and amplitudes over adapted projections, together with the unconstrained parameters ψ comprising the spectral projection P_z, utterance-level gate, and classification head. With injection dropout $b \sim \text{Bernoulli}(1-p)$, we minimize

$$\mathcal{L}(\theta) := \mathbb{E}(x,y) \sim \mathcal{D}\Big[(1-p)\ell_{\text{CE}}(f_\theta(x;1),y) + p\ell_{\text{CE}}(f_\theta(x;0),y)\Big], \tag{4}$$

subject to the corresponding manifold constraint $\mathcal{M}_\ell$ for every $\ell \in \mathcal{J}$, and $f_\theta(x;b) \in \mathbb{R}$ be the detector logit for waveform x under injection mask b. This jointly trains the spectral-injection and SSL-only states without inverted-dropout rescaling.

Riemannian updates. The amplitude vectors and unconstrained parameters receive ordinary Euclidean updates. The orthonormal factors are instead updated using Riemannian gradients and retractions (Edelman et al., 1998; Absil et al., 2008). For a constrained factor $X \in \text{St}_Q(d,s)$, define $P_Q^\perp := I_d - QQ^\top$ and $\text{sym}(A) := (A + A^\top)/2$. Given the Euclidean gradient $G = \nabla_X \mathcal{L}$, its tangent projection is $\xi = \Pi_X(G) := P_Q^\perp G - X\,\text{sym}(X^\top G)$. SILO then applies the polar retraction: $X^+ = R_X(-\eta\xi) = (X - \eta\xi)\left(I_s + \eta^2 \xi^\top \xi\right)^{-1/2}$, which preserves both orthonormality and orthogonality to the protected subspace. For the unrestricted factor V_c, we take Q empty, so $P_Q^\perp = I_d$.

4 VDD AND FVD: NEW-DOMAIN EVALUATION DATASETS

Existing ADD benchmarks cover important shifts in generators, languages, and recording conditions (Müller et al., 2022; Todisco et al., 2019; Yamagishi et al., 2021; Müller et al., 2024; Hou et al., 2024), but two practically important settings remain weakly represented: speech affected by voice disorders and recordings that last minutes rather than a few seconds. Therefore, we construct two *evaluation datasets*: Voice-Disorder Deepfake (VDD) and ForVoice Deepfake (FVD). Tab. 9 (Appendix) positions VDD and FVD against existing audio deepfake datasets.

4.1 DATASET STATISTICS

VDD: pathological-speech deepfake evaluation. VDD is derived from a Hungarian voice-disorder corpus (Jenei et al., 2021). All speakers read the same Hungarian passage, providing a controlled linguistic condition across five clinical groups: healthy controls, depression, Parkinson's disease, organic dysphonia, and functional dysphonia. Applying 7 generation systems to each source recording produces 4,942 spoof recordings, giving 5,648 evaluation trials and 84.8 h of audio in total. VDD evaluates a failure mode not captured by conventional ADD benchmarks: genuine pathological speech can exhibit irregular spectral and phonatory characteristics that a detector may incorrectly associate with spoofing, while synthesized speech may partially normalize these characteristics.

FVD: long-length deepfake evaluation. FVD is derived from the ForVoice120 forensic-speech corpus (Sztahó & Fejes, 2023). It contains 120 native Hungarian speakers, each recorded in two sessions and four elicitation tasks: free conversation, collaborative dialogue, narrative speech, and a fixed forensic phrase. Unlike conventional clip-level ADD datasets, these recordings range from tens of seconds to 12.3 minutes, with a median duration of 5.4 minutes. Applying the same 7 generation systems produces 6,720 full-length spoof recordings, resulting in 7,680 recording-level trials and approximately 666 h of audio. FVD enables evaluation of whether weak spoofing evidence distributed across individual segments can be accumulated into a reliable recording-level decision.

4.2 SPOOF GENERATION

Synthesis and voice-conversion systems. We use the same 7 generation systems for VDD and FVD, which include zero-shot voice cloning, autoregressive codec-language modeling, flow-matching synthesis, diffusion-based VC, and discrete-token VC. This broadens the evaluation beyond conventional vocoder artifacts. Particularly, 4 systems perform TTS: **A01: XTTS-v2** (Casanova et al., 2024), **A02: F5-TTS-Hungarian** (Chen et al., 2025; Maxdorger29, 2026), **A03: MMS-TTS-HUN** (Pratap et al., 2024), and **A04: Piper** (Rhasspy, 2023). The remaining 3 systems perform VC: **A05: MKL-VC** (Lobashev et al., 2025), **A06: Seed-VC** (Liu, 2024), and **A07: EZ-VC** (Joglekar et al., 2025). We store all final VDD and FVD trials as 16 kHz mono audio, and we handle reference speech selection and source duration independently for the two datasets, as described below.

VDD spoof generation. Spoofing is performed directly at the recording level. The common read passage provides the linguistic content for TTS, while reference-conditioned systems use speech from the corresponding source recording to preserve speaker characteristics. Voice-conversion systems operate on the bonafide speech using a separately selected target speaker. All generated recordings undergo the same final resampling and normalization procedure as the bonafide data to reduce preprocessing shortcuts. Construction details are provided in Appendix E.2.

FVD spoof generation. FVD requires a different construction because many source recordings are too long to synthesize in a single generation call. We therefore use a *chunk–generate–reassemble* pipeline based on the corpus-provided time-aligned annotations. Each recording is divided into speech segments, which are independently synthesized or converted and then placed back onto the original recording timeline. Reference-conditioned systems use speech from the same speaker in the other recording session, avoiding direct overlap between the conditioning and generated material. The reconstructed spoof preserves the duration and temporal organization of its bonafide counterpart, including the inter-segment structure, enabling controlled evaluation at the full-recording level. Detailed segmentation rules, reference selection, normalization, reconstruction, and generation-quality analyses are deferred to Appendix E.3.

5 EXPERIMENTS

Baseline Detectors and Datasets. We evaluate 13 detectors, including LCNN (Wu et al., 2020), RawNet2 (Tak et al., 2021), AASSIST (Jung et al., 2022), SpecCNN Müller et al. (2022), XLSR-Linear Babu et al. (2022), RawBMamba (Chen et al., 2024), IBA Wu et al. (2024), BiCrossMamba-ST (El Kheir et al., 2025), SSL+Mamba (Tran et al., 2025), AIsynth_voice (Ren et al., 2025), AASIST-ERM Laakkonen et al. (2026), AASIST-MLDG Laakkonen et al. (2026), and SONAR (Hidekel et al., 2026). All detectors are trained on WaveFake (Frank & Schönherr, 2021), and eveluated on ASVSpoof-19-LA (ASVS-19) (Todisco et al., 2019), ASVSpoof-21-DF (ASVS-21) (Yamagishi et al., 2021), In-the-Wild (ITW) (Müller et al., 2022), MLAAD-v3 (Müller et al., 2022), PolyGlotFake (Hou et al., 2024), CVoiceFake (Li et al., 2024a), our VDD and FVD datasets.

Metrics. We mainly report Equal Error Rate (EER), accuracy (ACC), and area under the ROC curve (AUC), all in percentage (%). Lower EER is better, while higher ACC and AUC are better. Unless otherwise stated, we use **macro-averaged metrics for MLAAD** since all audios in this dataset are grouped by generators. Implementation details are provided in Appendix D.

Table 1: **Cross-dataset audio deepfake detection performance**. Models are trained on WaveFake and evaluated on the in-domain WaveFake test set and four unseen cross-domain datasets.

Methods	Parameter	In-Domain			Cross-Domain															
		WaveFake			ASVSpoof-19			ASVSpoof-21			In-the-Wild (ITW)			MLAAD			Average		
		EER	ACC	AUC	EER	ACC	AUC	EER	ACC	AUC	EER	ACC	AUC	EER	ACC	AUC	EER	ACC	AUC
SpecCNN	0.58M	0.09	91.58	100	41.60	45.44	61.40	50.74	19.12	52.80	60.92	57.36	36.60	26.93	47.56	78.90	45.04	42.37	57.42
LCNN	0.16M	0.07	99.96	100	39.35	56.46	64.90	49.53	25.89	53.80	58.35	53.16	39.70	25.19	57.42	80.70	43.11	48.23	59.77
RawNet2	17.62M	0.368	99.75	99.99	37.34	23.87	64.30	45.10	40.51	55.90	54.19	35.68	42.90	28.71	64.93	75.70	41.33	41.25	59.70
AASIST	0.30M	0.161	99.78	100	40.86	45.88	64.90	37.64	47.72	68.70	26.09	66.72	81.90	38.37	60.89	65.60	35.74	55.30	70.27
XLSR Linear	315.57M	0.078	97.71	99.99	13.72	78.54	92.40	10.59	91.34	95.30	6.16	90.04	98.60	36.01	58.24	65.80	16.62	79.54	88.02
Aisynth_voice	29.63M	25.11	68.79	83.75	23.17	78.55	86.47	28.57	83.43	78.58	20.19	68.49	88.66	41.47	54.03	60.20	28.35	71.13	78.48
RawBMamba	0.72M	0.058	99.97	100	37.96	71.20	67.50	38.54	82.75	65.40	35.58	46.04	67.60	57.61	53.67	41.30	42.42	63.41	60.45
SSL + Mamba	317.52M	5.096	34.93	98.74	9.13	41.52	95.60	4.77	69.09	98.70	7.53	87.29	97.60	42.56	41.88	57.30	15.99	59.95	87.30
BiCrossMamba-ST	0.51M	0.046	99.88	100	34.03	83.68	74.30	35.30	93.70	71.70	38.70	38.02	63.90	47.24	60.70	54.50	38.81	69.03	66.10
IBA	52.32M	0.06	99.42	100	2.05	91.20	99.75	4.18	97.74	99.19	4.06	93.04	99.24	26.30	66.31	77.52	9.15	87.07	93.93
AASIST-ERM	3.59M	0	99.34	100	1.32	98.19	99.86	4.64	99.00	97.03	4.75	91.60	96.65	22.96	68.67	80.53	8.42	89.37	93.52
AASIST-MLDG	3.59M	0	98.85	100	1.82	97.86	99.76	2.14	98.99	99.02	2.70	92.83	98.55	23.04	67.86	80.69	7.43	89.39	94.51
SONAR	637.62M	0.004	99.86	100	1.57	96.78	99.90	1.78	99.24	99.90	5.10	88.56	97.90	20.49	74.71	82.80	7.23	90.24	95.12
SILO (Ours)	7.94M	0	99.98	100	1.02	99.02	99.90	1.56	99.39	99.70	4.35	89.79	97.50	14.38	78.59	89.50	**5.32**	**91.69**	**96.65**

5.1 PRIMARY RESULTS

5.1.1 CROSS-DATASET GENERALIZATION

Compare with SOTA Audio Deepfake Detectors. Results in Tab. 1 demonstrate two notable advantages of SILO. (1) **Generalizability:** SILO achieves the strongest overall cross-dataset performance among the compared detectors, indicating better robustness to unseen spoofing conditions rather than overfitting to the source dataset. The gains are particularly clear under challenging dataset shifts, showing that SILO captures more transferable fake-discriminative cues. (2) **Efficiency:** SILO obtains these improvements with a lightweight low-rank adaptation strategy, requiring only a small number of trainable parameters compared with large fully adapted detectors. This shows that strong cross-dataset generalization can be achieved without heavy parameter updates.

Table 2: **Cross-dataset comparison with LoRA-based methods (EER% ↓).** All methods use the XLS-R backbone Babu et al. (2021).

Methods	In-Domain	Cross-Domain			
	WaveFake	ASVS19	ASVS21	ITW	MLAAD
XLS-R+LoRA	0.01	<u>1.28</u>	<u>2.18</u>	**2.37**	26.83
Effort	5.88	23.46	15.94	12.01	27.19
XLS-R+StelLA	**0.00**	1.35	2.24	<u>2.50</u>	<u>21.24</u>
SILO	**0.00**	**1.02**	**1.56**	4.35	**14.38**

Compare with other LoRA Methods. We also compare SILO with other LoRA-based methods, including standard LoRA Hu et al. (2022), Effort (Yan et al., 2025), and Stella Li et al. (2026). For these methods, the SSL backbone is XLS-R. From Tab. 2, we can observe that SILO better balances robustness to different domain shifts compared to standard LoRA adaptation (e.g., XLR-R+LoRA and XLS-R+StelLA). Additionally, it is shown that directly transferring LoRA-based methods designed for the image domain to ADD can be ineffective: Effort substantially underperforms even conventional XLS-R+LoRA across all evaluated datasets.

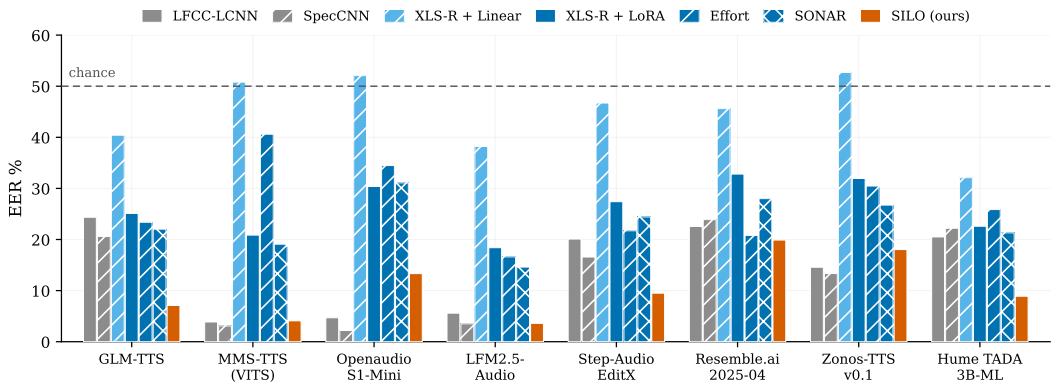


Figure 4: **Performance on modern unseen generators of MLAAD (EER% ↓)**. SILO maintains consistently low EER across diverse recent generators

Table 3: **Evaluation on VDD across healthy and pathological speaker groups**. All values are in %. EER ↓, ACC ↑, and TPR@1%FPR ↑.

	Healthy			Depression			Parkinson			Organic dysphonia			Functional dysphonia			Average		
Methods	**EER**	**ACC**	**TPR@1% FPR**	**EER**	**ACC**	**TPR@1% FPR**	**EER**	**ACC**	**TPR@1% FPR**	**EER**	**ACC**	**TPR@1% FPR**	**EER**	**ACC**	**TPR@1% FPR**	**EER**	**ACC**	**TPR@1% FPR**
SpecCNN	63.9	16.4	2.0	62.4	14.3	2.7	29.3	18.4	27.2	44.1	16.3	5.2	46.2	17.6	8.8	49.18	16.60	9.18
LCNN	60.2	24.8	1.9	58.5	22.9	3.1	28.1	30.7	19.7	47.5	25.6	7.7	46.0	28.5	11.0	48.06	26.50	8.68
RawNet2	48.7	84.6	1.3	39.9	85.6	1.1	43.4	86.0	2.4	54.3	83.0	0.2	47.9	84.6	0.0	46.84	84.76	1.00
AASIST	33.7	75.9	11.0	35.2	74.0	36.3	36.6	73.2	31.4	40.2	74.1	18.1	38.6	74.8	5.7	36.86	74.40	20.50
XLS-R Linear	26.8	68.0	25.8	38.6	63.6	16.0	31.9	69.3	17.6	38.8	65.7	8.3	30.1	65.8	24.8	33.24	66.48	18.50
Aisynth_voice	33.3	69.2	16.0	31.5	69.8	6.0	32.9	73.5	12.9	32.9	69.5	5.2	37.6	61.3	8.3	33.64	68.66	9.68
RawBMamba	29.2	86.7	0.0	29.9	86.8	0.0	37.2	87.8	0.0	35.0	85.5	0.0	38.0	85.5	0.0	33.86	86.46	0.00
SSL + Mamba	27.6	22.9	34.4	28.3	19.7	30.0	26.3	27.1	23.5	38.6	29.9	3.7	31.2	24.7	31.7	30.40	24.86	24.66
BiCrossMamba-ST	28.5	87.6	16.0	25.9	87.5	3.6	33.2	87.5	11.4	28.5	87.8	17.2	29.6	87.5	7.1	29.14	**87.58**	11.06
SONAR	15.3	83.4	74.2	16.1	81.1	10.1	28.3	83.8	5.5	31.9	79.2	18.4	25.3	79.8	12.0	23.38	81.46	24.04
IBA	25.96	66.47	49.67	32.16	63.61	35.78	30.88	67.58	12.79	38.48	69.83	12.92	29.89	65.86	11.60	31.474	66.67	24.55
AASIST-ERM	19.41	75.97	57.35	26.00	72.01	35.21	28.53	75.29	0.29	30.55	74.67	0.35	23.04	71.35	10.99	25.50	73.85	20.83
AASIST-MLDG	21.93	72.76	54.71	28.39	68.96	25.67	29.49	74.77	6.76	34.71	71.26	0.35	27.71	65.60	5.27	28.44	70.67	18.55
SILO (Ours)	14.6	78.3	79.4	16.9	87.2	60.1	20.9	83.5	38.2	27.9	75.8	30.3	22.8	78.5	43.7	**20.62**	80.66	**50.34**

5.1.2 Modern Generator Generalization

Fig. 4 shows that the classical spectral detectors remain highly effective for several novel generators, supporting our observation that spectral representations expose artifacts that transfer across synthesis families. By injecting spectral evidence, SILO consistently outperforms the SSL-based baselines across all evaluated generators and substantially reduces their failures on challenging synthesis systems. Moreover, SILO closes much of the gap to classical spectral detectors, showing that injecting spectral evidence successfully strengthens the SSL representation with synthesis-sensitive cues while retaining its robust backbone.

5.1.3 Evaluation on VDD and FVD Datasets

Tab. 3 reveals that **pathological speech remains a challenging distribution shift for existing ADD systems**. Notably, several detectors achieve relatively high ACC while exhibiting very low TPR@1%FPR. This discrepancy shows that aggregate accuracy can mask poor detection under a strict false-alarm constraint: when false alarms must be kept low, these detectors fail to identify a large fraction of deepfakes. Therefore, pathology-related speech characteristics substantially alter the separation between bonafide and spoof scores, exposing a failure mode that is not apparent from conventional healthy-speech evaluation. **SILO is consistently more robust** under this shift. Nevertheless, its performance also degrades markedly across pathological speaker groups and remains far from reliable. These results highlight the purpose of VDD: rather than providing another in-domain benchmark, it exposes pathological speech as an important and still unresolved generalization challenge for ADD.

We additionally evaluate all detectors on FVD to examine detection over long-form forensic recordings. Full FVD results across different recording durations are provided in Appendix F.1.

5.2 Ablation and Analysis

Effect of Constraints. Both constraints contribute to cross-domain generalization as shown in Tab. 4. Enforcing the Stiefel geometry with Riemannian optimization consistently improves over free adaptation, while additionally protecting the pretrained subspace further reduces EER on all four target datasets. The largest improvement appears on ITW (10.29% → 4.35%), suggesting that the two constraints are complementary in preserving transferable SSL representations while learning task-specific adaptations.

Table 4: **Effect of SILO constraints (EER% ↓).**

Setting	Protected	Stiefel	Optimizer	MLAAD	ITW	ASVS19	ASVS21
Free	No	No	AdamW	18.19	10.29	2.37	4.03
Unpreserved	No	Yes	Riemannian	17.16	7.63	2.00	2.59
Preserved	**Yes**	**Yes**	**Riemannian**	**14.38**	**4.35**	**1.02**	**1.56**

Table 5: **Effect of preserved energy π (EER% ↓).**

π	k	MLAAD	ITW	ASVS19	ASVS21
0	0	17.16	7.63	2.00	2.59
0.30	52	17.05	4.41	2.01	2.61
0.50	114	16.42	4.42	2.07	2.79
0.65	185	15.88	4.45	1.10	1.53
0.75	253	15.58	3.93	0.89	1.35
0.80	**295**	**14.38**	**4.35**	**1.02**	**1.56**
0.90	420	15.88	4.94	1.21	1.67
0.95	529	14.28	6.10	1.54	1.93
0.99	718	13.76	8.35	1.62	2.68

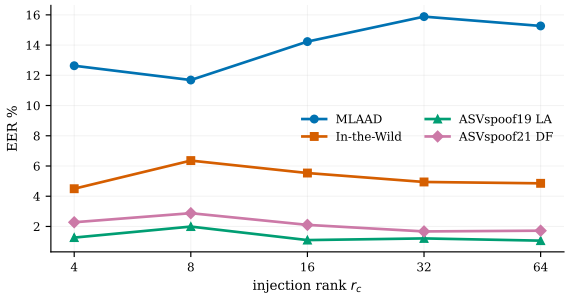


Figure 5: **Analysis of the spectral injection rank r_c.**

Tab. 5 further analyzes different π values. Performance is most favorable at intermediate preservation levels, confirming the need to balance pretrained knowledge preservation and adaptation capacity. Particularly, $\pi = 0.8$ provides a strong balance across the 4 target domains and is selected as our default setting. Increasing π beyond 0.8 progressively degrades ITW and ASVspoof performance, indicating that excessive preservation restricts the subspace available for cross-domain adaptation. Hyperparameters are selected on WaveFake's validation set, then re-evaluated on all test sets.

How much spectral-injection capacity is needed? Fig. 5 studies the different injection rank r_c, which controls the number of independent spectral directions available to the spectral branch. It shows that increasing the injection rank does not monotonically improve generalization. These results suggest that a moderate rank provides a useful compromise between capturing transferable spectral artifacts and avoiding excessive reliance on spectral cues. We therefore use $r_c = 32$ as the default configuration.

6 Conclusion

In this work, we propose **Spectral Injection via Low-rank Orthogonality (SILO)**, where utterance-adaptive spectral evidence complements a frozen SSL backbone through constrained low-rank adaptation. SILO combines a dedicated spectral-injection branch with geometric and preserved-subspace constraints that explicitly control adaptation strength while protecting selected high-energy pretrained directions, with these constraints maintained through optimization on restricted Stiefel manifolds. Extensive cross-dataset evaluations show that SILO improves generalization across unseen synthesis systems, including both classical and recent generators, while maintaining strong robustness under diverse domain shifts and acoustic conditions.

We also introduce **Voice-Disorder Deepfake (VDD)** and **ForVoice Deepfake (FVD)** to evaluate two underexplored settings. Results on VDD show that pathological speech can substantially disrupt the separation between bonafide and spoof speech, with several detectors performing poorly under strict false-positive constraints. FVD further reveals that while strong detectors benefit from the first 1–2 minutes of audio, performance can deteriorate beyond 2 minutes. This non-monotonic behavior shows that simply providing more audio does not guarantee better detection and highlights the need for more effective long-range evidence aggregation in ADD. We hope that SILO, together with VDD and FVD, provides both a stronger generalization framework and new evaluation settings for studying these remaining challenges toward practical and reliable ADD.

AI USE STATEMENT

We used generative AI tools in two capacities. First, to simulate real-world attack scenarios, generative speech models were used to construct the spoof portions of the VDD and FVD evaluation datasets, as described in Sec. 4.2 and Appendix E. Second, during the development of research code, large language models (LLMs) were moderately used under the authors' validation and verification to assist the implementation, debugging, and code refactoring. They were also used to support manuscript drafting and editing, focusing on fixing typo and grammar errors.

All AI-assisted content was reviewed and verified by the authors. AI-assisted code was inspected, executed, and tested by the authors before being used in the reported experiments. The authors take full responsibility for the final content of this work, including its text, claims, proofs, code, and experimental results.

ETHICS STATEMENT

This work uses existing speech corpora to construct the VDD and FVD evaluation datasets and does not involve the collection of new human-subject data. For both source corpora, we completed the consent and data-use procedures required by the respective dataset authors/providers and obtained permission to use the data for research purposes. We use these data only within the scope of the granted research permissions. In addition, we do not redistribute the original source recordings or the derived VDD and FVD datasets; accessing to the underlying recordings remains governed by the terms of the original data providers.

REPRODUCIBILITY STATEMENT

We provide detailed information to support reproduction of the proposed method and experiments throughout the paper and appendix. The SILO formulation, constraints, and optimization procedure are specified in Sec. 3, with notation summarized in Appendix B and complete proofs of the theoretical results provided in Appendix C. Appendix D reports the model architecture, initialization, optimization, training hyperparameters, and other implementation settings. The construction of VDD and FVD, including spoof-generation procedures, segmentation, reference selection, reconstruction, normalization, and generation-quality diagnostics, is described in Sec. 4 and Appendix E. Additional experimental results and analyses are provided in Appendix F.

REFERENCES

P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2008.

Mohd Mujtaba Akhtar, Muskaan Singh, et al. Hcfd: A benchmark for audio deepfake detection in healthcare. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 34829–34843, 2026.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick Von Platen, Yatharth Saraf, Juan Pino, et al. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*, 2021.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. Xls-r: Self-supervised cross-lingual speech representation learning at scale. In *Interspeech 2022*, pp. 2278–2282. ISCA, September 2022. doi: 10.21437/interspeech.2022-143. URL http://dx.doi.org/10.21437/interspeech.2022-143.

Paul Boersma et al. Accurate short-term analysis of the fundamental frequency and the harmonics-to-noise ratio of a sampled sound. In *Proceedings of the institute of phonetic sciences*, volume 17, pp. 97–110. Amsterdam, 1993.

Nicolas Boumal. An introduction to optimization on smooth manifolds. *To appear with Cambridge University Press,(Apr 2022)*, 2020.

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, and Julian Weber. Xtts: a massively multilingual zero-shot text-to-speech model. In *Interspeech 2024*, pp. 4978–4982. ISCA, September 2024. URL 10.21437/interspeech.2024-2016.

Yujie Chen, Jiangyan Yi, Jun Xue, Chenglong Wang, Xiaohui Zhang, Shunbo Dong, Siding Zeng, Jianhua Tao, Zhao Lv, and Cunhang Fan. Rawbmamba: End-to-end bidirectional state space model for audio deepfake detection. In *Interspeech 2024*, pp. 2720–2724. ISCA, September 2024. doi: 10.21437/interspeech.2024-698. URL http://dx.doi.org/10.21437/interspeech.2024-698.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, JianZhao JianZhao, Kai Yu, and Xie Chen. F5-tts: A fairytaler that fakes fluent and faithful speech with flow matching. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6255–6271, 2025.

Ariel Cohen, Inbal Rimon, Eran Aflalo, and Haim H Permuter. A study on data augmentation in voice anti-spoofing. *Speech Communication*, 141:56–67, 2022.

Anh-Tuan Dao, Nicholas Evans, and Driss Matrouf. Spoofing detection in the wild: an investigation of approaches to improve generalisation. In *Proc. Odyssey 2024*, pp. 145–150, 2024.

Steven Davis and Paul Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE transactions on acoustics, speech, and signal processing*, 28(4):357–366, 1980.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. In *Interspeech 2020*, pp. 3830–3834. ISCA, October 2020. doi: 10.21437/interspeech.2020-2650. URL http://dx.doi.org/10.21437/interspeech.2020-2650.

Alan Edelman, Tomás A Arias, and Steven T Smith. The geometry of algorithms with orthogonality constraints. *SIAM journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

Yassine El Kheir, Tim Polzehl, and Sebastian Möller. BiCrossMamba-ST: Speech Deepfake Detection with Bidirectional Mamba Spectro-Temporal Cross-Attention. In *Interspeech 2025*, pp. 2235–2239, 2025. doi: 10.21437/Interspeech.2025-527.

Joel Frank and Lea Schönherr. Wavefake: A data set to facilitate audio deepfake detection. *arXiv preprint arXiv:2111.02813*, 2021.

Gene H Golub and Charles F Van Loan. *Matrix computations*, volume 3. JHU press, 2013.

Ido Nitzan Hidekel, Gal Lifshitz, Khen Cohen, and Dan Raviv. SONAR: Spectral-contrastive audio residuals for generalizable deepfake detection. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=Jr7k590VxW.

Yang Hou, Haitao Fu, Chunkai Chen, Zida Li, Haoyu Zhang, and Jianjun Zhao. Polyglotfake: A novel multilingual and multimodal deepfake dataset. In *International conference on pattern recognition*, pp. 180–193. Springer, 2024.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM transactions on audio, speech, and language processing*, 29:3451–3460, 2021.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Wen Huang, Yanmei Gu, Zhiming Wang, Huijia Zhu, and Yanmin Qian. Generalizable audio deepfake detection via latent space refinement and augmentation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025a.

Wen Huang, Yanmei Gu, Zhiming Wang, Huijia Zhu, and Yanmin Qian. Speechfake: A large-scale multilingual speech deepfake dataset incorporating cutting-edge generation methods. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9985–9998, 2025b.

Attila Zoltán Jenei, Gábor Kiss, Miklós Gábriel Tulics, and Dávid Sztahó. Separation of several illnesses using correlation structures with convolutional neural networks. *Acta Polytechnica Hungarica*, 18(7):47–66, 2021.

Advait Joglekar, Divyanshu Singh, Rooshil Rohit Bhatia, and Srinivasan Umesh. Ez-vc: Easy zero-shot any-to-any voice conversion. In *EMNLP (Findings)*, pp. 19768–19774, 2025.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. *URL https://kellerjordan. github. io/posts/muon*, 6(3):4, 2024.

Jee-weon Jung, Hee-Soo Heo, Hemlata Tak, Hye-jin Shim, Joon Son Chung, Bong-Jin Lee, Ha-Jin Yu, and Nicholas Evans. Aasist: Audio anti-spoofing using integrated spectro-temporal graph attention networks. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 6367–6371. IEEE, 2022.

Atharva Kulkarni, Hoan My Tran, Ajinkya Kulkarni, Sandipana Dowerah, Damien Lolive, and Mathew Magimai Doss. Exploring generalization to unseen audio data for spoofing: Insights from ssl models. In *ASVSpoof workshop 2024*, 2024.

Janne Laakkonen, Ivan Kukanov, and Ville Hautamäki. Mixture of low-rank adapter experts in generalizable audio deepfake detection. In *2025 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 2211–2216. IEEE, 2025.

Janne Laakkonen, Ivan Kukanov, and Ville Hautamäki. Generalizable speech deepfake detection via meta-learned lora. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 19092–19096. IEEE, 2026.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

John M Lee. Smooth manifolds. In *Introduction to smooth manifolds*, pp. 1–31. Springer, 2013.

Xinfeng Li, Kai Li, Yifan Zheng, Chen Yan, Xiaoyu Ji, and Wenyuan Xu. Safeear: Content privacy-preserving audio deepfake detection. In *Proceedings of the 2024 on ACM SIGSAC Conference on Computer and Communications Security*, pp. 3585–3599, 2024a.

Yuang Li, Min Zhang, Mengxin Ren, Xiaosong Qiao, Miaomiao Ma, Daimeng Wei, and Hao Yang. Cross-domain audio deepfake detection: Dataset and analysis. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4977–4983, 2024b.

Zhizhong Li, Sina Sajadmanesh, Jingtao Li, and Lingjuan Lyu. Stella: Subspace learning in low-rank adaptation using stiefel manifold. *Advances in Neural Information Processing Systems*, 38: 75066–75092, 2026.

Kai Lion, Liang Zhang, Bingcong Li, and Niao He. Polar: Polar-decomposed low-rank adapter representation. *Advances in Neural Information Processing Systems*, 38:149313–149358, 2026.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, et al. Muon is scalable for llm training. *arXiv preprint arXiv:2502.16982*, 2025.

Songting Liu. Zero-shot voice conversion with diffusion transformers. *arXiv preprint arXiv:2411.09943*, 2024.

Alexander Lobashev, Assel Yermekova, and Maria Larchenko. Training-free voice conversion with factorized optimal transport. In *Interspeech 2025*, pp. 1373–1377. ISCA, 2025. URL http://dx.doi.org/10.21437/interspeech.2025-2043.

Maxdorger29. F5-TTS Hungarian: Native magyar text-to-speech with zero-shot voice cloning. Hugging Face model repository, 2026. URL https://huggingface.co/Maxdorger29/f5-tts-hungarian.

Nicolas M Müller, Pavel Czempin, Franziska Dieckmann, Adam Froghyar, and Konstantin Böttinger. Does audio deepfake detection generalize? *arXiv preprint arXiv:2203.16263*, 2022.

Nicolas M Müller, Piotr Kawa, Wei Herng Choong, Edresson Casanova, Eren Gölge, Thorsten Müller, Piotr Syga, Philip Sperl, and Konstantin Böttinger. Mlaad: The multi-language audio anti-spoofing dataset. In *2024 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–7. IEEE, 2024.

Olivia Murton, Robert Hillman, and Daryush Mehta. Cepstral peak prominence values for clinical voice evaluation. *American Journal of Speech-Language Pathology*, 29(3):1596–1607, 2020.

Hai-Son Nguyen-Le, Hung-Cuong Nguyen-Thanh, Nhien-An Le-Khac, Dinh-Thuc Nguyen, and Hong-Hanh Nguyen-Le. Afss: Artifact-focused self-synthesis for mitigating bias in audio deepfake detection. *arXiv preprint arXiv:2603.26856*, 2026.

Juneyoung Park, Minjae Kang, Seongbae Lee, Haegang Lee, Seongwan Kim, and Jaeho Lee. Riemannian optimization for lora on the stiefel manifold. In *EMNLP (Findings)*, pp. 20971–20985, 2025.

Octavian Pascu, Adriana Stan, Dan Oneata, Elisabeta Oneata, and Horia Cucu. Towards generalisable and calibrated audio deepfake detection with self-supervised representations. In *Proc. Interspeech 2024*, pp. 4828–4832, 2024.

Kaare Brandt Petersen, Michael Syskind Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 7(15):510, 2008.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52, 2024.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pp. 28492–28518. PMLR, 2023.

Hainan Ren, Li Lin, Chun-Hao Liu, Xin Wang, and Shu Hu. Improving generalization for ai-synthesized voice detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 20165–20173, 2025.

Rhasspy. Piper: A fast, local neural text to speech system. GitHub repository, 2023. URL https://github.com/rhasspy/piper.

Md Sahidullah, Tomi Kinnunen, and Cemal Hanilçi. A comparison of features for synthetic speech detection. 2015.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*, 2019.

Dávid Sztahó and Attila Fejes. Effects of language mismatch in automatic forensic voice comparison using deep learning embeddings. *Journal of forensic sciences*, 68(3):871–883, 2023.

Hemlata Tak, Jose Patino, Massimiliano Todisco, Andreas Nautsch, Nicholas Evans, and Anthony Larcher. End-to-end anti-spoofing with rawnet2. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6369–6373. IEEE, 2021.

Hemlata Tak, Massimiliano Todisco, Xin Wang, Jee-weon Jung, Junichi Yamagishi, and Nicholas Evans. Automatic speaker verification spoofing and deepfake detection using wav2vec 2.0 and data augmentation. *arXiv preprint arXiv:2202.12233*, 2022.

Massimiliano Todisco, Héctor Delgado, and Nicholas Evans. A new feature for automatic speaker verification anti-spoofing: Constant q cepstral coefficients. In *Proc. Odyssey 2016*, pp. 283–290, 2016.

Massimiliano Todisco, Xin Wang, Ville Vestman, Md. Sahidullah, Héctor Delgado, Andreas Nautsch, Junichi Yamagishi, Nicholas Evans, Tomi H. Kinnunen, and Kong Aik Lee. Asvspoof 2019: Future horizons in spoofed and fake audio detection. In *Interspeech 2019*, pp. 1008–1012. ISCA, 2019. URL 10.21437/interspeech.2019-2249.

Hoan My Tran, Damien Lolive, David Guennec, Aghilas Sini, Arnaud Delhay, and Pierre-François Marteau. Leveraging ssl speech features and mamba for enhanced deepfake detection. In *Interspeech 2025*, pp. 5323–5327, 2025.

Chenglong Wang, Jiangyan Yi, Xiaohui Zhang, Jianhua Tao, Le Xu, and Ruibo Fu. Low-rank adaptation method for wav2vec2-based fake audio detection. *IJCAI 2023 Workshop on Deepfake Audio Detection and Analysis (DADA)*, 2023.

Xiaopeng Wang, Ruibo Fu, Zhengqi Wen, Zhiyong Wang, Yuankun Xie, Yukun Liu, Jianhua Tao, Xuefei Liu, Yongwei Li, Xin Qi, et al. Genuine-focused learning using mask autoencoder for generalized fake audio detection. *arXiv preprint arXiv:2406.03247*, 2024a.

Xin Wang and Junichi Yamagishi. Can large-scale vocoded spoofed data improve speech spoofing countermeasure with a self-supervised front end? In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 10311–10315. IEEE, 2024.

Zhiyong Wang, Ruibo Fu, Zhengqi Wen, Yuankun Xie, Yukun Liu, Xiaopeng Wang, Xuefei Liu, Yongwei Li, Jianhua Tao, Yi Lu, et al. Generalized fake audio detection via deep stable learning. *arXiv preprint arXiv:2406.03237*, 2024b.

Haochen Wu, Wu Guo, Shengyu Peng, Zhuhai Li, and Jie Zhang. Adapter learning from pre-trained model for robust spoof speech detection. In *Interspeech*, 2024.

Zhenzong Wu, Rohan Kumar Das, Jichen Yang, and Haizhou Li. Light convolutional neural network with feature genuinization for detection of synthetic speech attacks. In *Interspeech 2020*, pp. 1101–1105. ISCA, October 2020. doi: 10.21437/interspeech.2020-1810. URL http://dx.doi.org/10.21437/interspeech.2020-1810.

Junichi Yamagishi, Xin Wang, Massimiliano Todisco, Md Sahidullah, Jose Patino, Andreas Nautsch, Xuechen Liu, Kong Aik Lee, Tomi Kinnunen, Nicholas Evans, et al. Asvspoof 2021: accelerating progress in spoofed and deepfake speech detection. *arXiv preprint arXiv:2109.00537*, 2021.

Zhiyuan Yan, Jiangming Wang, Peng Jin, Ke-Yue Zhang, Chengchun Liu, Shen Chen, Taiping Yao, Shouhong Ding, Baoyuan Wu, and Li Yuan. Orthogonal subspace decomposition for generalizable AI-generated image detection. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 70268–70288. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/yan25b.html.

Yuan Zhang, Jiang Hu, Zhijian Lai, Lin Lin, and Zaiwen Wen. Retraction-free optimization over the stiefel manifold for the lora fine-tuning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026.

Appendix roadmap. This appendix provides the supporting material for SILO, progressing from methodological context and theoretical foundations to implementation details and extended empirical evaluation. Appendix A discusses additional related work on generalized audio deepfake detection, geometric constraints for LoRA, and the distinction between SILO and Muon optimization. Appendix B summarizes the notation used throughout the paper. Appendix C provides the complete theoretical analysis, including the geometry of the restricted Stiefel manifold, the spectral properties and protected-subspace preservation of the proposed adaptation, the correctness of the restricted-Stiefel update, and the theoretical analysis of the spectral gate. Appendix D reports the implementation and training details of SILO together with the audio transformations used for robustness evaluation. Appendix E provides additional information on the VDD and FVD datasets, including their construction and characteristics. Appendix F presents extended experiments on FVD and VDD, robustness to audio transformations, generalization to unseen generators and languages, and additional ablation analyses. Finally, Appendix G discusses the limitations of the current study.

A Extended Related Work

A.1 Generalized Audio Deepfake Detection

Self-supervised representations. A prominent direction exploits self-supervised learning (SSL) models (Schneider et al., 2019; Babu et al., 2021; Hsu et al., 2021) pretrained on large and diverse speech corpora. Tak et al. (2022) combine XLS-R with an AASIST back-end and waveform augmentation, demonstrating the effectiveness of pretrained SSL representations for spoofing detection. Pascu et al. (2024) further show that frozen SSL representations coupled with lightweight classifiers can provide strong cross-dataset performance. However, strong SSL representations do not eliminate distribution shift. Kulkarni et al. (2024) report substantial performance variation across unseen datasets even with SSL-based detectors and investigate multi-stage training and model ensembles to improve generalization. These findings suggest that SSL representations provide a strong transferable basis for ADD, but may still encode insufficient or domain-dependent evidence for previously unseen synthesis processes.

Augmentation and generalization-oriented learning. Another research direction explicitly increases training diversity or modifies the learning objective to discourage reliance on domain-specific artifacts. Dao et al. (2024) investigate augmentation, pretrained-model fine-tuning, and Siamese comparison with bonafide reference speech for in-the-wild generalization. Cross-domain and attack-augmented training similarly increases the diversity of spoofing conditions observed during learning (Li et al., 2024b). Beyond augmentation, Wang et al. (2024b) introduce sample-weight-based stable learning to reduce feature correlations associated with distribution shift, whereas genuine-focused learning places greater emphasis on modeling bonafide speech rather than the heterogeneous spoof class (Wang et al., 2024a). Instead, Huang et al. (2025a) regularize the spoof embedding space using multiple prototypes together with latent-space augmentation. These approaches improve generalization by diversifying the training distribution or constraining which representations and decision boundaries are learned.

Parameter-efficient adaptation. Recent works have adopted parameter-efficient fine-tuning (PEFT) as a way to adapt large pretrained SSL models for ADD while reducing the number of trainable parameters. Adapter-based SSL fine-tuning utilizes Low-Rank Adaptation (LoRA) (Hu et al., 2022) to update forgery-relevant features only (Wu et al., 2024). Meta-learned LoRA explicitly learns low-rank adaptations that transfer across spoofing attacks (Laakkonen et al., 2026), while mixtures of LoRA experts specialize multiple low-rank adapters and route inputs among them for improved out-of-domain detection (Laakkonen et al., 2025). These approaches primarily address how pretrained SSL representations should be adapted efficiently for better transfer.

SILO is complementary to these generalization strategies. Rather than relying only on augmentation, a stronger SSL representation, or a more transferable adaptation rule, it explicitly introduces an additional source of spectral evidence into the pretrained SSL representation. Motivated by the failure modes of spectral and SSL representations, SILO uses a separate low-rank spectral-injection operator modulated by an utterance-level gate while constraining learned updates relative to selected singular subspaces of the frozen projection. Therefore, it addresses both *what complementary evidence is available during adaptation* and *where that evidence is permitted to modify the pretrained representation*.

A.2 Geometric Constraints for LoRA

Standard LoRA (Hu et al., 2022) represents a residual as $\Delta W = BA$, constraining its rank but not the geometry of its factors. Several recent methods introduce Stiefel-manifold constraints to improve the geometry and optimization of LoRA. Therefore, we do *not* claim that orthogonal LoRA factors, Stiefel-manifold optimization, or Riemannian updates are themselves novel.

LoRA with Stiefel manifold. Park et al. (2025) constrain one LoRA factor to the Stiefel manifold and optimize it with a Riemannian optimizer, reducing basis redundancy in standard LoRA and maintaining effective rank during fine-tuning. PoLAR (Lion et al., 2026) parameterizes the residual as $\Delta W = X\Theta Y^\top$, where both X and Y have orthonormal columns and Θ is unconstrained. Similarly, StelLA (Li et al., 2026) uses $\Delta W = USV^\top$ and optimizes both U and V on Stiefel manifolds, separating the learned input/output subspaces from the middle scaling matrix. Retraction-free Manifold-LoRA (Zhang et al., 2026) instead focuses on optimization, avoiding explicit retractions through a landing method for Stiefel-constrained LoRA. These prior works have already explored Stiefel-manifold constraints, three-factor LoRA parameterizations, and Riemannian optimization.

Difference from SILO. SILO also uses orthonormal direction factors, but its main geometric distinction is that its feasible subspaces depend explicitly on the frozen pretrained weight. Given $W_0 = \bar{U}\,\mathrm{diag}(\sigma)\bar{V}^\top$, SILO constrains the learned directions to be orthogonal to selected high-energy singular subspaces of W_0. In contrast, prior methods such as PoLAR and StelLA enforce orthogonality *within* the learned LoRA factors. SILO additionally enforces orthogonality *between* the learned factors and designated pretrained subspaces. This restricted Stiefel geometry yields the protected-subspace preservation property established in Theorem **??**. Another distinction is that SILO applies this geometry constraint to a two-source ADD architecture rather than to a single conventional LoRA residual. The SSL-adaptation branch adapts the pretrained representation, while the spectral-injection branch injects an external, time-aligned spectral representation under an utterance-dependent gate. Therefore, in comparison with existing geometric LoRA methods, SILO combines three ingredients: (i) directly readable residual spectra through signed diagonal amplitudes; (ii) pretrained-weight-dependent restricted Stiefel constraints with exact protected-subspace invariance; and (iii) a separate gated low-rank operator for injecting complementary spectral evidence. The first shares substantial motivation with existing geometric parameterizations, whereas the latter two define the main methodological distinction of SILO.

A.3 Comparison with Muon Optimization

Muon (Liu et al., 2025; Jordan et al., 2024) also exploits matrix orthogonalization, but at the *optimizer* level. Given a momentum matrix, Muon applies Newton–Schulz iterations to approximately orthogonalize it and uses the resulting matrix as the update direction. Thus, orthogonality is imposed on the *update*, not on the model parameters themselves. In contrast, SILO parameterizes each low-rank residual as $U\,\mathrm{diag}(\alpha)V^\top$ and constrains its direction factors to restricted Stiefel manifolds defined relative to the frozen projection W_0. Riemannian projection and retraction maintain these constraints throughout training, thereby preserving selected pretrained singular subspaces. Hence, Muon uses orthogonalization to improve optimization, whereas SILO uses manifold optimization to enforce a structural constraint on the learned adaptation.

B Summary of Notation

Tab. 6 summarizes the notation used in the main paper. Symbols introduced only inside a single proof are defined where they first appear in Appendix C.

C Proofs

C.1 Additional Notation

Appendix B summarizes the notation of the main paper, while this section fixes the additional notations used in the proofs.

Table 6: Summary of the notation used in this paper.

Symbol	Meaning
Inputs and dimensions	
x	Input waveform
d	Width of the SSL hidden state and of the frozen projection
d_s	Output width of the spectral encoder
$h_t \in \mathbb{R}^d$	SSL hidden representation at step t
$z_t \in \mathbb{R}^d$	Time-aligned spectral feature at step t
$E_{\text{spec}},\ P_z$	Frozen spectral encoder and its learned projection to $\mathbb{R}^d$
Frozen projection and preserved subspaces	
$W_0 \in \mathbb{R}^{d\times d}$	Frozen square projection
$\bar{U}\,\mathrm{diag}(\sigma)\bar{V}^\top$	SVD of W_0, with $\sigma_1 \geq \cdots \geq \sigma_d \geq 0$
$\pi \in [0, 1)$	Preserved-energy fraction
k_π, abbreviated k	Preserved rank induced by π
$U_k, V_k \in \mathbb{R}^{d\times k}$	Preserved left and right singular bases of W_0
$U_\perp, V_\perp \in \mathbb{R}^{d\times(d-k)}$	Remaining left and right singular vectors, $\bar{U} = [\,U_k\ U_\perp\,]$, $\bar{V} = [\,V_k\ V_\perp\,]$
$\Sigma_k,\ \Sigma_\perp$	$\mathrm{diag}(\sigma_1, \ldots, \sigma_k)$ and $\mathrm{diag}(\sigma_{k+1}, \ldots, \sigma_d)$
Residual maps and ranks	
r	*Residual rank* of the SSL-adaptation branch
r_c	*Injection rank*, i.e. the bandwidth of the spectral-injection branch
$U, V \in \mathbb{R}^{d\times r}$, $\alpha \in \mathbb{R}^r$	Factors and signed amplitudes of the SSL residual
$U_c, V_c \in \mathbb{R}^{d\times r_c}$, $\alpha_c \in \mathbb{R}^{r_c}$	Factors and signed amplitudes of the spectral injection map
$\Delta W_{\text{ssl}} = U\,\mathrm{diag}(\alpha)V^\top$	SSL residual map
$\Delta W_c = U_c\,\mathrm{diag}(\alpha_c)V_c^\top$	Spectral injection map
$W_{\text{adapt}} := W_0 + \Delta W_{\text{ssl}}$	Adapted SSL projection
$\mathcal{F}_\beta(h_t, z_t)$	SILO layer output at gate value β
$\mathcal{R}_\beta = [\,\Delta W_{\text{ssl}}\ \ \beta\Delta W_c\,]$	Two-input residual operator $\mathbb{R}^d \oplus \mathbb{R}^d \to \mathbb{R}^d$
Gating and injection dropout	
$g(x) \in (0, 1)$	Utterance-level spectral gate
$b \in \{0, 1\}$	Injection mask, drawn once per utterance
p	Injection-dropout probability, $b \sim \mathrm{Bernoulli}(1-p)$
$\beta(x) = b\,g(x)$	Effective gate; β also denotes a fixed gate value
Constraints and optimization	
$\mathrm{St}(d, s)$	Stiefel manifold $\{X \in \mathbb{R}^{d\times s} : X^\top X = I_s\}$
$\mathrm{St}_Q(d, s)$	Restricted Stiefel manifold, additionally imposing $Q^\top X = 0$
$\mathcal{M}$	Product manifold of feasible factors at one projection
$\Pi_X(G)$	Projection of the Euclidean gradient G onto $T_X\mathrm{St}_Q(d, s)$
$R_X(-\eta\xi)$	Polar retraction with step size η and $\xi := \Pi_X(G)$

For matrices A and B of the same size, write $\langle A, B\rangle_F := \mathrm{tr}(A^\top B)$ for the Frobenius inner product. For a square matrix M, define $\mathrm{sym}(M) := (M + M^\top)/2$ and $\mathrm{skew}(M) := (M - M^\top)/2$. Let $\mathrm{Sym}(s)$ be the vector space of real symmetric $s \times s$ matrices. If $Q \in \mathbb{R}^{d\times k}$ has orthonormal columns, then $Q_\perp \in \mathbb{R}^{d\times(d-k)}$ denotes a fixed matrix whose columns form an orthonormal basis of $\mathrm{range}(Q)^\perp$, and $P_Q^\perp := I_d - QQ^\top = Q_\perp Q_\perp^\top$. When $k = 0$, Q is the empty $d \times 0$ matrix, $QQ^\top = 0$, and $\mathrm{St}_Q(d, s) = \mathrm{St}(d, s)$. For a positive-definite matrix H, $H^{-1/2}$ denotes its principal inverse square root.

For $W_0 = \bar{U}\,\mathrm{diag}(\sigma)\bar{V}^\top$, we write $\bar{U} = [\,U_k\ U_\perp\,]$ and $\bar{V} = [\,V_k\ V_\perp\,]$, where the complementary matrices consist of the remaining singular vectors, and we set $\Sigma_k := \mathrm{diag}(\sigma_1, \ldots, \sigma_k)$ and $\Sigma_\perp :=$

$\mathrm{diag}(\sigma_{k+1}, \ldots, \sigma_d)$, so that $W_0 = U_k \Sigma_k V_k^\top + U_\perp \Sigma_\perp V_\perp^\top$. Then

$$P_U^\perp := I_d - U_k U_k^\top = U_\perp U_\perp^\top, \qquad P_V^\perp := I_d - V_k V_k^\top = V_\perp V_\perp^\top. \tag{5}$$

C.2 Feasibility and geometry of the restricted Stiefel manifold

The ordinary Stiefel manifold $\mathrm{St}(n, s) = \{X \in \mathbb{R}^{n \times s} : X^\top X = I_s\}$ has a standard smooth-manifold structure (Edelman et al., 1998; Absil et al., 2008; Boumal, 2020). The additional constraint $Q^\top X = 0$ simply restricts the columns of X to the orthogonal complement of a fixed subspace. The following lemmas make this reduction precise and provide the geometric properties used in Sec. 3.

Lemma 1 (Feasibility of the restricted Stiefel manifold)**.** *Let $d \geq 1$ and $0 \leq k < d$ be integers, let $Q \in \mathbb{R}^{d \times k}$ satisfy $Q^\top Q = I_k$, let $Q_\perp \in \mathbb{R}^{d \times (d-k)}$ be a fixed matrix whose columns form an orthonormal basis of $\mathrm{range}(Q)^\perp$, and let $1 \leq s \leq d$ be an integer. Then:*

(i) $\mathrm{St}_Q(d, s) \neq \varnothing$ if and only if $s \leq d - k$;

(ii) If $s \leq d - k$, then the map $\Phi : \mathrm{St}(d - k, s) \to \mathrm{St}_Q(d, s)$, $\Phi(\widetilde{X}) := Q_\perp \widetilde{X}$, is a bijective Frobenius isometry, with inverse $\Phi^{-1}(X) = Q_\perp^\top X$. In particular, $X \in \mathrm{St}_Q(d, s) \iff X = Q_\perp \widetilde{X}$ for a unique $\widetilde{X} \in \mathrm{St}(d - k, s)$.

Proof. Since the columns of Q form an orthonormal basis of $\mathrm{range}(Q)$ and the columns of $Q_\perp$ form an orthonormal basis of $\mathrm{range}(Q)^\perp$, the square matrix $[\, Q \; Q_\perp \,]$ is orthogonal. Consequently,

$$\underbrace{Q^\top Q_\perp = 0}_{(A)}, \qquad \underbrace{Q_\perp^\top Q_\perp = I_{d-k}}_{(B)}, \qquad \underbrace{I_d = QQ^\top + Q_\perp Q_\perp^\top}_{(C)}. \tag{6}$$

For (i), first suppose that $\mathrm{St}_Q(d, s) \neq \varnothing$ and choose $X \in \mathrm{St}_Q(d, s)$. Since $X^\top X = I_s$, the columns of X are orthonormal, and hence $\mathrm{rank}(X) = s$. Moreover, $Q^\top X = 0$ implies $QQ^\top X = Q(Q^\top X) = 0$. Applying the term (C) in Eq. 6 gives

$$X = I_d X = (QQ^\top + Q_\perp Q_\perp^\top) X = Q_\perp Q_\perp^\top X. \tag{7}$$

Thus, every column of X lies in $\mathrm{range}(Q_\perp) = \mathrm{range}(Q)^\perp$. Since $Q_\perp$ has $d - k$ orthonormal columns, $s = \mathrm{rank}(X) \leq \mathrm{rank}(Q_\perp) = d - k$. Therefore, $s \leq d - k$ is necessary.

Conversely, suppose that $s \leq d - k$. Select any s columns of $Q_\perp$ and collect them in a matrix $X \in \mathbb{R}^{d \times s}$. These columns are orthonormal, so $X^\top X = I_s$. They also lie in $\mathrm{range}(Q)^\perp$, so $Q^\top X = 0$. Hence $X \in \mathrm{St}_Q(d, s)$, proving that $\mathrm{St}_Q(d, s)$ is nonempty. Therefore,

$$\boxed{\mathrm{St}_Q(d, s) \neq \varnothing \iff s \leq d - k.} \tag{8}$$

This proves part (i).

For (ii), assume $s \leq d - k$ and define $\Phi : \mathrm{St}(d - k, s) \longrightarrow \mathrm{St}_Q(d, s),\ \Phi(\widetilde{X}) = Q_\perp \widetilde{X}$.

First, for $\widetilde{X} \in \mathrm{St}(d - k, s)$, Eq. 6 yields:

$$Q^\top \Phi(\widetilde{X}) = Q^\top Q_\perp \widetilde{X} = 0, \tag{9}$$

$$\Phi(\widetilde{X})^\top \Phi(\widetilde{X}) = \widetilde{X}^\top Q_\perp^\top Q_\perp \widetilde{X} = \widetilde{X}^\top \widetilde{X} = I_s. \tag{10}$$

Therefore, $\Phi(\widetilde{X}) \in \mathrm{St}_Q(d, s)$.

Now let $X \in \mathrm{St}_Q(d, s)$ and define $\widetilde{X} := Q_\perp^\top X$. Since $Q^\top X = 0$, Eq. 7 gives:

$$X = Q_\perp Q_\perp^\top X = Q_\perp \widetilde{X} = \Phi(\widetilde{X}). \tag{11}$$

Moreover, we have $\widetilde{X}^\top \widetilde{X} = X^\top Q_\perp Q_\perp^\top X = X^\top X = I_s$, where the second equality follows from $Q_\perp Q_\perp^\top X = X$. Hence $\widetilde{X} \in \mathrm{St}(d - k, s)$, and Φ is **surjective**.

For every $\widetilde{X} \in \mathrm{St}(d-k, s)$, $Q_\perp^\top \Phi(\widetilde{X}) = Q_\perp^\top Q_\perp \widetilde{X} = \widetilde{X}$. Thus, Φ is **injective**. Together with Eq. 11, this also gives $\Phi^{-1}(X) = Q_\perp^\top X$. In particular, every $X \in \mathrm{St}_Q(d, s)$ has the unique representation $X = Q_\perp \widetilde{X}$ with $\widetilde{X} \in \mathrm{St}(d-k, s)$.

Finally, for $\widetilde{X}_1, \widetilde{X}_2 \in \mathrm{St}(d-k, s)$, we have:

$$\begin{aligned} \Big\langle \Phi(\widetilde{X}_1), \Phi(\widetilde{X}_2) \Big\rangle_F &= \mathrm{tr}\Big((Q_\perp \widetilde{X}_1)^\top Q_\perp \widetilde{X}_2 \Big) \\ &= \mathrm{tr}\Big(\widetilde{X}_1^\top Q_\perp^\top Q_\perp \widetilde{X}_2 \Big) \\ &= \mathrm{tr}(\widetilde{X}_1^\top \widetilde{X}_2) \\ &= \langle \widetilde{X}_1, \widetilde{X}_2 \rangle_F. \end{aligned} \tag{12}$$

Therefore, Φ preserves the Frobenius inner product and the associated norm and distance. This proves part (ii). □

Lemma 2 (Manifold structure and tangent space of $\mathrm{St}_Q(d, s)$)**.** *Let d, k, and Q be as in Lemma 1, and let $1 \le s \le d-k$ be an integer. Then $\mathrm{St}_Q(d, s)$ is a compact embedded submanifold of $\mathbb{R}^{d \times s}$ of dimension $(d-k)s - \frac{1}{2}s(s+1)$, and for every $X \in \mathrm{St}_Q(d, s)$ its tangent space is*

$$T_X \,\mathrm{St}_Q(d, s) = \{ Z \in \mathbb{R}^{d \times s} : Q^\top Z = 0,\ \mathrm{sym}(X^\top Z) = 0 \}. \tag{13}$$

Proof. We represent the two constraints of $\mathrm{St}_Q(d, s)$ by the smooth map:

$$\Psi : \mathbb{R}^{d \times s} \longrightarrow \mathbb{R}^{k \times s} \times \mathrm{Sym}(s), \qquad \Psi(Y) := (Q^\top Y,\ Y^\top Y - I_s). \tag{14}$$

A matrix Y belongs to $\mathrm{St}_Q(d, s)$ exactly when both components of $\Psi(Y)$ are zero. Thus, $\mathrm{St}_Q(d, s) = \Psi^{-1}(0, 0)$.

We next compute the differential of Ψ. Let $X \in \mathrm{St}_Q(d, s)$ and let $Z \in \mathbb{R}^{d \times s}$.

- For the first component, $Q^\top (X + tZ) = Q^\top X + tQ^\top Z$, so its derivative at $t = 0$ is $Q^\top Z$.
- For the second component, $(X + tZ)^\top (X + tZ) - I_s = X^\top X - I_s + t(X^\top Z + Z^\top X) + t^2 Z^\top Z$.

Therefore, $D\Psi_X[Z] = \big(Q^\top Z,\ X^\top Z + Z^\top X \big)$.

To apply the regular level set theorem (Lee, 2013), we show that $D\Psi_X$ is surjective. Let $B \in \mathbb{R}^{k \times s}$ and $S \in \mathrm{Sym}(s)$ be arbitrary. Choose $Z = QB + \frac{1}{2} XS$.

Since $Q^\top Q = I_k$ and $Q^\top X = 0$,

$$Q^\top Z = Q^\top QB + \frac{1}{2} Q^\top XS = B. \tag{15}$$

Since $X^\top Q = (Q^\top X)^\top = 0$, $X^\top X = I_s$, and $S^\top = S$,

$$\begin{aligned} X^\top Z + Z^\top X &= X^\top \left(QB + \frac{1}{2} XS \right) + \left(QB + \frac{1}{2} XS \right)^\top X \\ &= X^\top QB + \frac{1}{2} X^\top XS + B^\top Q^\top X + \frac{1}{2} S^\top X^\top X \\ &= \frac{1}{2} S + \frac{1}{2} S = S. \end{aligned} \tag{16}$$

Hence: $D\Psi_X \big[QB + \frac{1}{2} XS \big] = (B, S)$.

Because (B, S) is arbitrary, $D\Psi_X$ is surjective at every $X \in \Psi^{-1}(0, 0)$. Thus, $(0, 0)$ is a regular value of Ψ. The regular-level-set theorem now implies that $\mathrm{St}_Q(d, s)$ is an embedded submanifold of $\mathbb{R}^{d \times s}$.

The ambient space $\mathbb{R}^{d\times s}$ has dimension ds. The codomain of Ψ has dimension:

$$\dim \mathbb{R}^{k\times s} + \dim \mathrm{Sym}(s) = ks + \frac{s(s+1)}{2}. \tag{17}$$

Since the differential is surjective, these constraints are independent. Therefore, the dimension of the level set is:

$$\boxed{\dim \mathrm{St}_Q(d,s) = ds - ks - \frac{s(s+1)}{2} = (d-k)s - \frac{s(s+1)}{2}.} \tag{18}$$

We next prove compactness. The map Ψ is continuous and $\{(0,0)\}$ is closed, so $\mathrm{St}_Q(d,s) = \Psi^{-1}(0,0)$ is closed. In addition, every $X \in \mathrm{St}_Q(d,s)$ satisfies $\|X\|_F^2 = \mathrm{tr}(X^\top X) = \mathrm{tr}(I_s) = s$. Thus, the manifold is bounded. Since it is a closed and bounded subset of the finite-dimensional Euclidean space $\mathbb{R}^{d\times s}$, it is compact.

It remains to identify the tangent space. For a regular level set, the tangent space at X is the kernel of the differential of the defining map. Hence:

$$T_X \,\mathrm{St}_Q(d,s) = \ker D\Psi_X = \left\{ Z \in \mathbb{R}^{d\times s} : Q^\top Z = 0,\ X^\top Z + Z^\top X = 0 \right\}. \tag{19}$$

By the definition of the symmetric part, $X^\top Z + Z^\top X = 2\,\mathrm{sym}(X^\top Z)$. Therefore, the second condition is equivalent to $\mathrm{sym}(X^\top Z) = 0$, and

$$\boxed{T_X \,\mathrm{St}_Q(d,s) = \left\{ Z \in \mathbb{R}^{d\times s} : Q^\top Z = 0,\ \mathrm{sym}(X^\top Z) = 0 \right\}.} \tag{20}$$

This proves the stated manifold, dimension, compactness, and tangent-space claims. □

C.3 Proof of Proposition 1

Proof. Under Constraint C1, the factor matrices satisfy $U^\top U = V^\top V = I_r,\ U_c^\top U_c = V_c^\top V_c = I_{r_c}$.

We first determine the singular values of the SSL-adaptation residual $\Delta W_{\mathrm{ssl}} = U D_\alpha V^\top$. Using $U^\top U = I_r$ and the fact that D_α is diagonal, we obtain:

$$\Delta W_{\mathrm{ssl}}^\top \Delta W_{\mathrm{ssl}} = (U D_\alpha V^\top)^\top (U D_\alpha V^\top) = V D_\alpha U^\top U D_\alpha V^\top = V D_\alpha^2 V^\top. \tag{21}$$

Let v_i be the ith column of V. Since the columns of V are orthonormal, $V^\top v_i = e_i$, where e_i denotes the ith standard basis vector of $\mathbb{R}^r$. Hence, $\Delta W_{\mathrm{ssl}}^\top \Delta W_{\mathrm{ssl}} v_i = V D_\alpha^2 V^\top v_i = \alpha_i^2 v_i,\ i = 1,\dots,r$. This shows that v_i is an eigenvector of $\Delta W_{\mathrm{ssl}}^\top \Delta W_{\mathrm{ssl}}$ with eigenvalue α_i^2.

Next, let $x \in \mathrm{range}(V)^\perp$. Orthogonality to every column of V gives $V^\top x = 0$, and hence:

$$\Delta W_{\mathrm{ssl}}^\top \Delta W_{\mathrm{ssl}} x = V D_\alpha^2 V^\top x = 0. \tag{22}$$

The r columns of V, together with any orthonormal basis of $\mathrm{range}(V)^\perp$, form an orthonormal basis of $\mathbb{R}^d$. Consequently, the eigenvalues of $\Delta W_{\mathrm{ssl}}^\top \Delta W_{\mathrm{ssl}}$, counted with multiplicity, are $\alpha_1^2, \dots, \alpha_r^2$ and $d-r$ additional zeros.

The singular values of a matrix are the nonnegative square roots of the eigenvalues of its Gram matrix. Therefore, the singular values of ΔW_{ssl} are:

$$|\alpha_1|, \dots, |\alpha_r| \quad \text{and} \quad d-r \text{ additional zeros.} \tag{23}$$

The same principle applies to $\Delta W_c = U_c D_{\alpha_c} V_c^\top$. Indeed, $U_c^\top U_c = I_{r_c}$ gives $\Delta W_c^\top \Delta W_c = V_c D_{\alpha_c}^2 V_c^\top$. Hence, the singular values of ΔW_c are:

$$|\alpha_{c,1}|, \dots, |\alpha_{c,r_c}| \quad \text{and} \quad d-r_c \text{ additional zeros.} \tag{24}$$

This proves part (i).

For $1 \le q < \infty$, the Schatten-q norm is the ℓ_q norm of the singular-value vector. The zero singular values do not contribute, so

$$\|\Delta W_{\mathrm{ssl}}\|_{S_q} = \left(\sum_{i=1}^{r} |\alpha_i|^q \right)^{1/q} = \|\alpha\|_q, \tag{25}$$

$$\|\Delta W_c\|_{S_q} = \left(\sum_{i=1}^{r_c} |\alpha_{c,i}|^q \right)^{1/q} = \|\alpha_c\|_q. \tag{26}$$

For $q = \infty$, the Schatten-∞ norm is the largest singular value. Therefore,

$$\boxed{\|\Delta W_{\text{ssl}}\|_{S_\infty} = \|\alpha\|_\infty, \qquad \|\Delta W_c\|_{S_\infty} = \|\alpha_c\|_\infty.} \tag{27}$$

Together, these identities prove $\|\Delta W_{\text{ssl}}\|_{S_q} = \|\alpha\|_q$, and $\|\Delta W_c\|_{S_q} = \|\alpha_c\|_q$ for every $1 \le q \le \infty$.

It remains to prove part (iii), the stated pointwise bound for the two-input residual operator and its operator-norm consequence. Let $\beta \in \mathbb{R}$ and $h, z \in \mathbb{R}^d$. By definition, $\mathcal{R}_\beta(h, z) = \Delta W_{\text{ssl}} h + \beta \Delta W_c z$. Applying the triangle inequality gives:

$$\begin{aligned} \|\mathcal{R}_\beta(h,z)\|_2 &= \|\Delta W_{\text{ssl}} h + \beta \Delta W_c z\|_2 \\ &\le \|\Delta W_{\text{ssl}} h\|_2 + \|\beta \Delta W_c z\|_2 \\ &= \|\Delta W_{\text{ssl}} h\|_2 + |\beta| \, \|\Delta W_c z\|_2. \end{aligned} \tag{28}$$

For any matrix A and vector x, the definition of the spectral norm gives $\|Ax\|_2 \le \|A\|_2 \|x\|_2$. Applying this inequality to the two residual matrices yields:

$$\|\mathcal{R}_\beta(h,z)\|_2 \le \|\Delta W_{\text{ssl}}\|_2 \|h\|_2 + |\beta| \, \|\Delta W_c\|_2 \|z\|_2. \tag{29}$$

The spectral norm is the Schatten-∞ norm. By the identities already proved,

$$\|\Delta W_{\text{ssl}}\|_2 = \|\alpha\|_\infty, \qquad \|\Delta W_c\|_2 = \|\alpha_c\|_\infty. \tag{30}$$

Substitution gives

$$\boxed{\|\mathcal{R}_\beta(h,z)\|_2 \le \|\alpha\|_\infty \|h\|_2 + |\beta| \, \|\alpha_c\|_\infty \|z\|_2,} \tag{31}$$

which is exactly the residual-norm bound stated in part (iii) of Proposition 1. Finally, applying the Cauchy–Schwarz inequality in $\mathbb{R}^2$ to the vectors $(\|\alpha\|_\infty, |\beta| \, \|\alpha_c\|_\infty)$ and $(\|h\|_2, \|z\|_2)$ gives

$$\|\mathcal{R}_\beta(h,z)\|_2 \le \left(\|\alpha\|_\infty^2 + \beta^2 \|\alpha_c\|_\infty^2\right)^{1/2} \left(\|h\|_2^2 + \|z\|_2^2\right)^{1/2} = \left(\|\alpha\|_\infty^2 + \beta^2 \|\alpha_c\|_\infty^2\right)^{1/2} \|(h,z)\|, \tag{32}$$

and taking the supremum over $\|(h, z)\| \le 1$ yields the stated bound on the operator norm of $\mathcal{R}_\beta$. This proves part (iii). □

C.4 Proof of Theorem 1

Proof. We first derive the consequences of the preserved-subspace constraint that will be used throughout the proof. Since $\Delta W_{\text{ssl}} = U D_\alpha V^\top$ and $U_k^\top U = 0$, $U_k^\top \Delta W_{\text{ssl}} = U_k^\top U D_\alpha V^\top = 0$.

Likewise, $V_k^\top V = 0$ implies $V^\top V_k = 0$, and therefore $\Delta W_{\text{ssl}} V_k = U D_\alpha V^\top V_k = 0$.

For the spectral-injection branch, $\Delta W_c = U_c D_{\alpha_c} V_c^\top$ and $U_k^\top U_c = 0$ give $U_k^\top \Delta W_c = U_k^\top U_c D_{\alpha_c} V_c^\top = 0$.

Taking the transpose of $U_k^\top \Delta W_{\text{ssl}} = 0$ also gives $\Delta W_{\text{ssl}}^\top U_k = 0$. Thus, the SSL residual cannot write into the protected output subspace or modify inputs from the protected SSL-input subspace, while the spectral-injection residual cannot write into the protected output subspace.

The partitioned SVD of the frozen projection is:

$$W_0 = [\, U_k \; U_\perp \,] \begin{bmatrix} \Sigma_k & 0 \\ 0 & \Sigma_\perp \end{bmatrix} [\, V_k \; V_\perp \,]^\top = U_k \Sigma_k V_k^\top + U_\perp \Sigma_\perp V_\perp^\top. \tag{33}$$

Because $[\, U_k \; U_\perp \,]$ and $[\, V_k \; V_\perp \,]$ are orthogonal,

$$U_k^\top U_\perp = 0, \qquad V_k^\top V_\perp = 0, \qquad U_k^\top U_k = V_k^\top V_k = I_k. \tag{34}$$

It follows directly that $U_k^\top W_0 = \Sigma_k V_k^\top, W_0 V_k = U_k \Sigma_k, W_0^\top U_k = V_k \Sigma_k$.

For (i), recall that the SILO projection at the fixed gate value β is:

$$\mathcal{F}_\beta(h,z) = W_0 h + \Delta W_{\text{ssl}} h + \beta \Delta W_c z. \tag{35}$$

For arbitrary $h, z \in \mathbb{R}^d$ and $\beta \in \mathbb{R}$, projecting its output onto the protected left singular subspace gives:

$$\boxed{\begin{aligned} U_k^\top \mathcal{F}_\beta(h,z) &= U_k^\top W_0 h + U_k^\top \Delta W_{\text{ssl}} h + \beta U_k^\top \Delta W_c z \\ &= \Sigma_k V_k^\top h + 0 + 0 = \Sigma_k V_k^\top h = U_k^\top W_0 h. \quad \text{(by C2)} \end{aligned}} \tag{36}$$

The protected output coordinates are therefore exactly the same as those produced by the frozen projection W_0. This proves (i).

For (ii), let $v \in \mathrm{range}(V_k)$. Then $v = V_k a$ for some $a \in \mathbb{R}^k$. Using $\Delta W_{\mathrm{ssl}} V_k = 0$, we obtain:

$$W_{\mathrm{adapt}} v = (W_0 + \Delta W_{\mathrm{ssl}}) V_k a = W_0 V_k a + \Delta W_{\mathrm{ssl}} V_k a = W_0 V_k a = \boxed{W_0 v}. \tag{37}$$

Hence, the adapted SSL projection agrees exactly with the frozen projection on $\mathrm{range}(V_k)$. This proves part (ii).

For (iii), let e_i denote the ith standard basis vector in $\mathbb{R}^k$. For every $i \le k$,

$$\bar{v}_i = V_k e_i, \qquad \bar{u}_i = U_k e_i, \qquad \Sigma_k e_i = \sigma_i e_i. \tag{38}$$

By part (ii), $\Delta W_{\mathrm{ssl}} V_k = 0$, we obtain:

$$\begin{aligned} W_{\mathrm{adapt}} \bar{v}_i &= (W_0 + \Delta W_{\mathrm{ssl}}) V_k e_i \\ &= W_0 V_k e_i \\ &= U_k \Sigma_k e_i \\ &= \boxed{\sigma_i \bar{u}_i}. \end{aligned} \tag{39}$$

Moreover, transposing $U_k^\top \Delta W_{\mathrm{ssl}} = 0$ gives $\Delta W_{\mathrm{ssl}}^\top U_k = 0$. Hence,

$$\begin{aligned} W_{\mathrm{adapt}}^\top \bar{u}_i &= (W_0^\top + \Delta W_{\mathrm{ssl}}^\top) U_k e_i \\ &= W_0^\top U_k e_i \\ &= V_k \Sigma_k e_i \\ &= \boxed{\sigma_i \bar{v}_i}. \end{aligned} \tag{40}$$

Since $\sigma_i \ge 0$, these are the left and right singular-vector relations associated with the singular value σ_i. Therefore, $(\sigma_i, \bar{u}_i, \bar{v}_i)$ remains an exact singular triplet of W_{adapt} for every $i \le k$. This proves part (iii).

For (iv), expand W_{adapt} in the pretrained left and right singular-vector bases:

$$\begin{bmatrix} U_k^\top \\ U_\perp^\top \end{bmatrix} W_{\mathrm{adapt}} \begin{bmatrix} V_k & V_\perp \end{bmatrix} = \begin{bmatrix} U_k^\top W_{\mathrm{adapt}} V_k & U_k^\top W_{\mathrm{adapt}} V_\perp \\ U_\perp^\top W_{\mathrm{adapt}} V_k & U_\perp^\top W_{\mathrm{adapt}} V_\perp \end{bmatrix}. \tag{41}$$

We evaluate the four components separately. For the upper-left component, we have:

$$U_k^\top W_{\mathrm{adapt}} V_k = U_k^\top W_0 V_k + U_k^\top \Delta W_{\mathrm{ssl}} V_k = \Sigma_k. \tag{42}$$

For the upper-right component, the partitioned SVD gives $U_k^\top W_0 V_\perp = 0$, while $U_k^\top \Delta W_{\mathrm{ssl}} = 0$ gives $U_k^\top \Delta W_{\mathrm{ssl}} V_\perp = 0$. Hence, $U_k^\top W_{\mathrm{adapt}} V_\perp = 0$.

For the lower-left component, the partitioned SVD gives $U_\perp^\top W_0 V_k = 0$, while $\Delta W_{\mathrm{ssl}} V_k = 0$ gives $U_\perp^\top \Delta W_{\mathrm{ssl}} V_k = 0$. Thus, $U_\perp^\top W_{\mathrm{adapt}} V_k = 0$.

Finally, for the lower-right component,

$$U_\perp^\top W_{\mathrm{adapt}} V_\perp = U_\perp^\top W_0 V_\perp + U_\perp^\top \Delta W_{\mathrm{ssl}} V_\perp = \Sigma_\perp + U_\perp^\top \Delta W_{\mathrm{ssl}} V_\perp. \tag{43}$$

Combining these four components yields:

$$\boxed{\begin{bmatrix} U_k^\top \\ U_\perp^\top \end{bmatrix} W_{\mathrm{adapt}} \begin{bmatrix} V_k & V_\perp \end{bmatrix} = \begin{bmatrix} \Sigma_k & 0 \\ 0 & \Sigma_\perp + U_\perp^\top \Delta W_{\mathrm{ssl}} V_\perp \end{bmatrix}}. \tag{44}$$

This proves part (iv) and completes the proof. □

C.5 CORRECTNESS OF THE RESTRICTED-STIEFEL UPDATE

Proposition 2 (Tangent projection and polar retraction on $\mathrm{St}_Q(d,s)$)**.** *Let d, k, and Q be as in Lemma 1, let $1 \le s \le d-k$ be an integer, and let $P_Q^\perp := I_d - QQ^\top$. For $X \in \mathrm{St}_Q(d,s)$ and $G \in \mathbb{R}^{d\times s}$, define $\Pi_X(G) := P_Q^\perp G - X\,\mathrm{sym}(X^\top G)$. Then:*

(i) For every $X \in \mathrm{St}_Q(d,s)$, the linear map $\Pi_X : \mathbb{R}^{d\times s} \to \mathbb{R}^{d\times s}$ is the orthogonal projection onto $T_X\,\mathrm{St}_Q(d,s)$ with respect to the Frobenius inner product.

(ii) For every $G \in \mathbb{R}^{d\times s}$ and every step size $\eta > 0$, setting $\xi := \Pi_X(G)$, the matrix $I_s + \eta^2\xi^\top\xi$ is symmetric positive definite and the update $X^+ := (X - \eta\xi)(I_s + \eta^2\xi^\top\xi)^{-1/2}$ satisfies $X^+ \in \mathrm{St}_Q(d,s)$.

(iii) For $X \in \mathrm{St}_Q(d,s)$ and $Z \in T_X\,\mathrm{St}_Q(d,s)$, the matrix $I_s + Z^\top Z$ is symmetric positive definite, and $R_X(Z) := (X+Z)(I_s+Z^\top Z)^{-1/2}$ lies in $\mathrm{St}_Q(d,s)$. The map $R : (X,Z) \mapsto R_X(Z)$ is a retraction on $\mathrm{St}_Q(d,s)$: it is smooth on the tangent bundle, $R_X(0) = X$, and $\frac{d}{dt}R_X(tZ)\big|_{t=0} = Z$ for every $Z \in T_X\,\mathrm{St}_Q(d,s)$; equivalently, $R_X(tZ) = X + tZ + O(t^2)$ as $t \to 0$, with the remainder measured in the Frobenius norm. Moreover, the update in (ii) is $X^+ = R_X(-\eta\xi)$.

Proof. By Lemma 2, $T_X\,\mathrm{St}_Q(d,s) = \{Z \in \mathbb{R}^{d\times s} : Q^\top Z = 0,\ X^\top Z + Z^\top X = 0\}$. Since $Q^\top X = 0$ and $X^\top X = I_s$, we also have $Q^\top P_Q^\perp = 0,\ X^\top P_Q^\perp = X^\top$.

For (i), we first show that $\Pi_X(G)$ is a tangent vector. Let $S := \mathrm{sym}(X^\top G)$, the defining identities give $Q^\top\Pi_X(G) = 0,$ and $X^\top\Pi_X(G) = X^\top G - S = \mathrm{skew}(X^\top G)$. The second expression is skew-symmetric, so $\Pi_X(G) \in T_X\,\mathrm{St}_Q(d,s)$.

To establish orthogonal projection, it remains to show that the residual $G - \Pi_X(G)$ is orthogonal to every tangent vector. We have $G - \Pi_X(G) = QQ^\top G + XS$. For any $Z \in T_X\,\mathrm{St}_Q(d,s)$,

$$\langle G - \Pi_X(G), Z\rangle_F = \mathrm{tr}(G^\top QQ^\top Z) + \mathrm{tr}(SX^\top Z) = 0. \tag{45}$$

The first term is zero because $Q^\top Z = 0$. For the second term, S is symmetric, whereas $X^\top Z$ is skew-symmetric, so their Frobenius inner product is zero. Hence $\langle G - \Pi_X(G), Z\rangle_F = 0$. Therefore, $\Pi_X(G)$ is the Frobenius-orthogonal projection of G onto $T_X\,\mathrm{St}_Q(d,s)$.

If $G = \nabla_X\mathcal{L}$, then, for every $Z \in T_X\,\mathrm{St}_Q(d,s)$,

$$D\mathcal{L}(X)[Z] = \langle G, Z\rangle_F = \langle \Pi_X(G), Z\rangle_F. \tag{46}$$

Since $\Pi_X(G)$ lies in the tangent space, it is the Riemannian gradient under the metric induced by the Frobenius inner product. This proves (i).

For (ii), part (i) gives $Q^\top\xi = 0$ and $X^\top\xi + \xi^\top X = 0$. Let $Y := X - \eta\xi$, then:

$$Q^\top Y = 0, \qquad Y^\top Y = I_s - \eta(X^\top\xi + \xi^\top X) + \eta^2\xi^\top\xi = I_s + \eta^2\xi^\top\xi. \tag{47}$$

Write $H := I_s + \eta^2\xi^\top\xi$. For every nonzero $a \in \mathbb{R}^s$, $a^\top Ha = \|a\|_2^2 + \eta^2\|\xi a\|_2^2 > 0$. Hence H is symmetric positive definite, and its principal inverse square root $H^{-1/2}$ exists and is symmetric. Since $X^+ = YH^{-1/2}$,

$$Q^\top X^+ = (Q^\top Y)H^{-1/2} = 0, \qquad (X^+)^\top X^+ = H^{-1/2}Y^\top YH^{-1/2} = H^{-1/2}HH^{-1/2} = I_s. \tag{48}$$

Both defining constraints hold, so $X^+ \in \mathrm{St}_Q(d,s)$. This proves (ii).

For (iii), let $Z \in T_X\,\mathrm{St}_Q(d,s)$. The tangent-space conditions give $Q^\top Z = 0,\ X^\top Z + Z^\top X = 0$. Consequently, we have:

$$Q^\top(X+Z) = 0, \qquad (X+Z)^\top(X+Z) = I_s + Z^\top Z =: H_Z. \tag{49}$$

The matrix H_Z is symmetric positive definite for every Z, so its principal inverse square root exists. Therefore, $R_X(Z) = (X+Z)H_Z^{-1/2}$ satisfies $Q^\top R_X(Z) = 0,$ and $R_X(Z)^\top R_X(Z) = H_Z^{-1/2}H_ZH_Z^{-1/2} = I_s$. Thus, R_X maps the entire tangent space into $\mathrm{St}_Q(d,s)$.

At $Z = 0$, $R_X(0) = X$. To examine the behavior near $Z = 0$, fix a tangent vector Z and replace it by tZ. Then:

$$R_X(tZ) = (X + tZ)(I_s + t^2 Z^\top Z)^{-1/2}. \tag{50}$$

Using the expansion $(I_s + t^2 Z^\top Z)^{-1/2} = I_s - \frac{t^2}{2} Z^\top Z + O(t^4)$, we obtain:

$$\begin{aligned} R_X(tZ) &= (X + tZ)\left(I_s - \frac{t^2}{2} Z^\top Z + O(t^4)\right) \\ &= X + tZ - \frac{t^2}{2} X Z^\top Z + O(t^3) = X + tZ + O(t^2). \end{aligned} \tag{51}$$

In particular,

$$\left.\frac{d}{dt} R_X(tZ)\right|_{t=0} = Z. \tag{52}$$

The inverse-square-root operation is smooth for symmetric positive-definite matrices, so $R_X(Z)$ depends smoothly on X and Z. Together with $R_X(0) = X$, $\left.\frac{d}{dt} R_X(tZ)\right|_{t=0} = Z$, these properties show that R_X is a retraction.

Finally, $X^+ = R_X(-\eta\xi)$, so the update in part (ii) is exactly the polar retraction applied to the projected descent direction. This proves (iii). When Q is empty, $QQ^\top = 0$ and the condition $Q^\top Z = 0$ disappears. The same formulas reduce to the standard tangent projection and polar retraction on $\mathrm{St}(d, s)$. □

C.6 Theoretical analysis of the spectral gate

We analyze two properties of SILO's utterance-level spectral gate. First, we characterize the gradient signal by which the detection objective learns the gate. Second, we quantify exactly how the gate scales injections entering through the spectral-injection branch. The first analysis follows from the Bernoulli negative log-likelihood and the multivariate chain rule used in backpropagation (LeCun et al., 2015). The second uses the induced Euclidean operator norm, its singular-value characterization (Golub & Van Loan, 2013), and standard quadratic-form/trace identities (Petersen et al., 2008).

C.6.1 What learning signal does the gate receive?

The gate construction, injection dropout, and training objective are defined in Secs. 3.3 and 3.5. We begin by examining the dependence of the detection loss on the shared injection strength. For a fixed training example (x, y), with $y \in \{0, 1\}$, let a denote the pre-sigmoid scalar produced by a_ϕ for x, and abbreviate $g := g_\phi(x) = \rho(a)$. We work at points where the derivatives with respect to a, g, and ϕ exist.

To isolate the effect of the gate value from the binary injection indicator in $f_\theta(x; b)$, we introduce an auxiliary logit $\widetilde{f}_\theta(x, \boldsymbol{\beta})$. For each adapted projection $\ell \in \mathcal{J}$, let β_ℓ denote an externally specified gate value, and write $\boldsymbol{\beta} = (\beta_\ell)_{\ell \in \mathcal{J}}$. The quantity $\widetilde{f}_\theta(x, \boldsymbol{\beta})$ is defined as the logit obtained when the gate at projection ℓ is fixed to β_ℓ. Because these gate values are specified externally, $\widetilde{f}_\theta$ does not depend on the gate parameters ϕ once $\boldsymbol{\beta}$ is fixed.

When all adapted projections use the same gate value β, define $\widehat{f}_\theta(x, \beta) := \widetilde{f}_\theta(x, \beta\mathbf{1})$. With this notation, $\widehat{f}_\theta(x, g) = f_\theta(x; 1)$, while $\widehat{f}_\theta(x, 0) = f_\theta(x; 0)$. The function $\widehat{f}_\theta$ still represents the complete nonlinear detector; in particular, no affine dependence of the logit on β is assumed.

For the fixed example and a fixed dropout probability $p \in [0, 1]$, the conditional loss is:

$$\ell_x(a) := \mathbb{E}_b[\ell_{\mathrm{CE}}(f_\theta(x; b), y)] = (1 - p)\ell_{\mathrm{CE}}\big(\widehat{f}_\theta(x, g), y\big) + p\ell_{\mathrm{CE}}\big(\widehat{f}_\theta(x, 0), y\big).$$

The second term uses the same jointly trained detector with spectral injection disabled. It is not the loss of a separately trained SSL-only model.

Proposition 3 (Task-dependent learning signal of the gate)**.** *Suppose that $\beta \mapsto \widehat{f}_\theta(x, \beta)$ is differentiable at $\beta = g$. Then the derivative of the dropout-averaged loss with respect to the gate logit a is*

$$\frac{\partial \ell_x}{\partial a} = (1 - p)g(1 - g)\left[\rho\Big(\widehat{f}_\theta(x, g)\Big) - y\right] \left.\frac{\partial \widehat{f}_\theta(x, \beta)}{\partial \beta}\right|_{\beta = g}. \tag{53}$$

If, in addition, $\phi \mapsto a$ is differentiable and the gate parameters ϕ affect ℓ_x only through a, then ℓ_x is differentiable with respect to ϕ, with

$$\nabla_\phi \ell_x = \frac{\partial \ell_x}{\partial a} \nabla_\phi a. \tag{54}$$

Proposition 3 shows that the gate is learned directly from the detection objective rather than from an auxiliary measure of spectral reliability. Its update is determined jointly by the detector's prediction error and the effect of the spectral-injection strength on the final logit. More precisely, when $p < 1$, the sign of the gate update is determined by the local derivative of the classification loss with respect to the shared injection strength: gradient descent increases the gate when an infinitesimal increase in spectral injection would decrease the loss, and decreases the gate when it would increase the loss. The factor $g(1-g)$ attenuates this learning signal when the sigmoid gate is close to saturation. The gate therefore provides a task-dependent mechanism for controlling how much spectral evidence is used for each utterance.

Proof. Fix (x, y), p, and all nongate parameters. For every scalar logit $t \in \mathbb{R}$, the binary cross-entropy loss is $\ell_{\mathrm{CE}}(t, y) = -y \log \rho(t) - (1-y) \log\big(1 - \rho(t)\big)$. Since $\rho'(t) = \rho(t)(1-\rho(t))$, differentiation gives:

$$\begin{aligned} \frac{\partial \ell_{\mathrm{CE}}(t,y)}{\partial t} &= -y \frac{\rho'(t)}{\rho(t)} + (1-y) \frac{\rho'(t)}{1-\rho(t)} = -y\big(1-\rho(t)\big) + (1-y)\rho(t) \\ &= \rho(t) - y. \end{aligned} \tag{55}$$

When $b = 0$, every adapted projection uses the externally fixed gate value zero. Consequently, $\widehat{f}_\theta(x, 0)$ is independent of a, and $\frac{\partial}{\partial a} \ell_{\mathrm{CE}}\big(\widehat{f}_\theta(x,0), y\big) = 0$. Differentiating the finite sum in Eq. 53 yields $\frac{\partial \ell_x}{\partial a} = (1-p) \frac{\partial}{\partial a} \ell_{\mathrm{CE}}\big(\widehat{f}_\theta(x,g), y\big)$. For the remaining term, the dependence on a is the composition $a \mapsto g = \rho(a) \mapsto \widehat{f}_\theta(x,g) \mapsto \ell_{\mathrm{CE}}(\widehat{f}_\theta(x,g), y)$. The scalar chain rule gives:

$$\begin{aligned} \frac{\partial \ell_x}{\partial a} &= (1-p) \left. \frac{\partial \ell_{\mathrm{CE}}(t,y)}{\partial t} \right|_{t = \widehat{f}_\theta(x,g)} \left. \frac{\partial \widehat{f}_\theta(x,\beta)}{\partial \beta} \right|_{\beta = g} \frac{\partial g}{\partial a} \\ &= (1-p) \Big[\rho\big(\widehat{f}_\theta(x,g)\big) - y\Big] \left. \frac{\partial \widehat{f}_\theta(x,\beta)}{\partial \beta} \right|_{\beta = g} g(1-g), \end{aligned}$$

Rearranging the scalar factors proves: $\boxed{\dfrac{\partial \ell_x}{\partial a} = (1-p)\, g(1-g) \left[\rho\Big(\hat{f}_\theta(x,g)\Big) - y\right] \left. \dfrac{\partial \hat{f}_\theta(x,\beta)}{\partial \beta} \right|_{\beta=g}}$.

For completeness, if $\boldsymbol{\beta} \mapsto \widetilde{f}_\theta(x, \boldsymbol{\beta})$ is also differentiable at $\boldsymbol{\beta} = g\mathbf{1}$, the definition of $\widehat{f}_\theta$ and the multivariate chain rule imply:

$$\left. \frac{\partial \widehat{f}_\theta(x,\beta)}{\partial \beta} \right|_{\beta=g} = \sum_{\ell \in \mathcal{J}} \left. \frac{\partial \widetilde{f}_\theta(x,\boldsymbol{\beta})}{\partial \beta_\ell} \right|_{\boldsymbol{\beta} = g\mathbf{1}}. \tag{56}$$

Thus, the derivative in equation 53 accounts for the gate's use at all adapted projections, rather than at a single projection.

Finally, if ϕ affects the loss only through a, the multivariate chain rule gives $\boxed{\nabla_\phi \ell_x = \dfrac{\partial \ell_x}{\partial a} \nabla_\phi a.}$ □

C.6.2 How does the gate control spectral injection?

We next characterize how the gate controls the contribution of the spectral branch to an adapted projection.

Consider one adapted projection with fixed parameters. For an externally specified gate value $\beta \in [0,1]$ and inputs $h, z \in \mathbb{R}^d$, the spectral-injection contribution is $\mathcal{F}_\beta(h,z) - \mathcal{F}_0(h,z) = \beta \Delta W_c z$, which does not depend on h. Therefore, we write $\mathcal{I}_\beta(z) := \beta \Delta W_c z$.

Proposition 4 (Gate control of spectral injection)**.** *Under Constraints C1 and C2, the following statements hold for fixed projection parameters and an externally specified gate value $\beta \in [0, 1]$.*

(i) For every spectral input $z \in \mathbb{R}^d$, $|\mathcal{I}_\beta(z)|2^2 = \beta^2 \sum_{i=1}^{r_c} \alpha_{c,i}^2 (v_{c,i}^\top z)^2$, where $v_{c,i}$ is the ith column of V_c. In particular, $|\mathcal{I}_\beta(z)|2 \le \beta |\alpha_c|\infty |z|_2$.

(ii) The largest possible spectral-injection gain is $\sup_{z\neq 0} \frac{\|\mathcal{I}_\beta(z)\|_2}{\|z\|_2} = \beta\|\alpha_c\|_\infty$. Moreover, for $0 \le \beta_1 \le \beta_2 \le 1$ with $\beta_2 > 0$, $\mathcal{I}_{\beta_1}(z) = \frac{\beta_1}{\beta_2}\mathcal{I}_{\beta_2}(z)$. Hence, $\|\mathcal{I}_{\beta_1}(z)\|_2 \le \|\mathcal{I}_{\beta_2}(z)\|_2$.

(iii) If z is random with finite second moment and $M_z := \mathbb{E}[zz^\top]$, then

$$\mathbb{E}\|\mathcal{I}_\beta(z)\|_2^2 = \beta^2 \sum_{i=1}^{r_c} \alpha_{c,i}^2 v_{c,i}^\top M_z v_{c,i} \le \beta^2 \|\alpha_c\|_\infty^2 \mathbb{E}\|z\|_2^2. \tag{57}$$

Thus, when the same fixed scalar β is applied to a fixed distribution of spectral inputs, the expected squared magnitude of the injected contribution scales as β^2.

(iv) The spectral-injection contribution has no component in the protected output subspace: $U_k^\top \mathcal{I}_\beta(z) = 0$ for every $z \in \mathbb{R}^d$.

Proposition 4 formalizes the control provided by the spectral gate. For fixed projection parameters and spectral input, decreasing the gate reduces the spectral-injection magnitude proportionally, while its expected squared magnitude decreases quadratically. Therefore, the gate provides direct control over how strongly spectral evidence modifies the layer output. This control does not affect the preservation guarantee: for every gate value, the injected contribution remains orthogonal to the protected output subspace.

Proof. The constraints give $U_c^\top U_c = V_c^\top V_c = I_{r_c}$, $U_k^\top U_c = 0$, and the spectral-injection map is $\Delta W_c = U_c \operatorname{diag}(\alpha_c) V_c^\top$. We prove the four claims in order.

(i) Using $U_c^\top U_c = I_{r_c}$ and the symmetry of the diagonal matrix $\operatorname{diag}(\alpha_c)$, we obtain

$$\begin{aligned}\|\mathcal{I}_\beta(z)\|_2^2 &= \beta^2 z^\top V_c \operatorname{diag}(\alpha_c) U_c^\top U_c \operatorname{diag}(\alpha_c) V_c^\top z \\ &= \beta^2 (V_c^\top z)^\top \operatorname{diag}(\alpha_c)^2 (V_c^\top z) = \beta^2 \sum_{i=1}^{r_c} \alpha_{c,i}^2 (v_{c,i}^\top z)^2.\end{aligned}$$

Since $V_c^\top V_c = I_{r_c}$, the matrix $V_c V_c^\top$ is the orthogonal projection onto $\operatorname{range}(V_c)$. Therefore, the orthogonal decomposition $z = V_c V_c^\top z + (I_d - V_c V_c^\top) z$ gives: $\|V_c^\top z\|_2^2 = \|V_c V_c^\top z\|_2^2 = \|z\|_2^2 - \|(I_d - V_c V_c^\top) z\|_2^2 \le \|z\|_2^2$. Together with $\alpha_{c,i}^2 \le \|\alpha_c\|_\infty^2$, this yields:

$$\begin{aligned}\|\mathcal{I}_\beta(z)\|_2^2 &\le \beta^2 \|\alpha_c\|_\infty^2 \sum_{i=1}^{r_c} (v_{c,i}^\top z)^2 = \beta^2 \|\alpha_c\|_\infty^2 \|V_c^\top z\|_2^2 \\ &\le \beta^2 \|\alpha_c\|_\infty^2 \|z\|_2^2.\end{aligned} \tag{58}$$

Taking nonnegative square roots and using $\beta \ge 0$, we conclude that: $\boxed{\|\mathcal{I}_\beta(z)\|_2 \le \beta \|\alpha_c\|_\infty \|z\|_2}$.

(ii) Dividing the bound in (i) by $\|z\|_2$ for $z \neq 0$ gives $\sup_{z\neq 0} \frac{\|\mathcal{I}_\beta(z)\|_2}{\|z\|_2} \le \beta \|\alpha_c\|_\infty$. To prove equality, choose an index j satisfying $|\alpha_{c,j}| = \|\alpha_c\|_\infty$ and set $z = v_{c,j}$. The orthonormality of the columns of V_c gives $\|v_{c,j}\|_2 = 1$, and all terms in the identity from (i) vanish except the term indexed by j. Hence:

$$\|\mathcal{I}_\beta(v_{c,j})\|_2^2 = \beta^2 \alpha_{c,j}^2, \qquad \frac{\|\mathcal{I}_\beta(v_{c,j})\|_2}{\|v_{c,j}\|_2} = \beta \|\alpha_c\|_\infty.$$

Therefore, the upper bound is attained, proving $\boxed{\sup_{z\neq 0} \frac{\|\mathcal{I}_\beta(z)\|_2}{\|z\|_2} = \beta \|\alpha_c\|_\infty}$. This argument also covers $\beta = 0$ and $\alpha_c = 0$.

For $0 \le \beta_1 \le \beta_2 \le 1$ with $\beta_2 > 0$, the definition of $\mathcal{I}_\beta$ gives

$$\mathcal{I}_{\beta_1}(z) = \beta_1 \Delta W_c z = \frac{\beta_1}{\beta_2}\big(\beta_2 \Delta W_c z\big) = \frac{\beta_1}{\beta_2}\mathcal{I}_{\beta_2}(z).$$

Since $\beta_1/\beta_2 \in [0,1]$, homogeneity of the Euclidean norm implies $\|\mathcal{I}_{\beta_1}(z)\|_2 = \frac{\beta_1}{\beta_2}\|\mathcal{I}_{\beta_2}(z)\|_2 \le \|\mathcal{I}_{\beta_2}(z)\|_2$.

(iii) Suppose that $\mathbb{E}\|z\|_2^2 < \infty$. Each column $v_{c,i}$ has unit norm, so the Cauchy–Schwarz inequality gives $(v_{c,i}^\top z)^2 \le \|z\|_2^2$. In particular, every term in the finite sum in (i) is integrable. Since β and the projection parameters are fixed, linearity of expectation gives:

$$\mathbb{E}\|\mathcal{I}_\beta(z)\|_2^2 = \beta^2 \sum_{i=1}^{r_c} \alpha_{c,i}^2 \mathbb{E}\big[(v_{c,i}^\top z)^2\big]. \tag{59}$$

For each i, the definition of M_z yields:

$$\mathbb{E}\big[(v_{c,i}^\top z)^2\big] = \mathbb{E}\big[v_{c,i}^\top z z^\top v_{c,i}\big] = v_{c,i}^\top \mathbb{E}[zz^\top] v_{c,i} = v_{c,i}^\top M_z v_{c,i}. \tag{60}$$

Substituting this identity proves:

$$\boxed{\mathbb{E}\|\mathcal{I}_\beta(z)\|_2^2 = \beta^2 \sum_{i=1}^{r_c} \alpha_{c,i}^2 v_{c,i}^\top M_z v_{c,i}.} \tag{61}$$

Taking expectations in the squared bound from (i) also gives: $\boxed{\mathbb{E}\|\mathcal{I}_\beta(z)\|_2^2 \le \beta^2 \|\alpha_c\|_\infty^2 \mathbb{E}\|z\|_2^2.}$

For fixed projection parameters and a fixed input distribution, the sum in the exact identity is independent of β, establishing the stated quadratic scaling. No zero-mean assumption on z is required, since M_z is its second-moment matrix, not necessarily its covariance matrix.

(iv) The protected-subspace constraint $U_k^\top U_c = 0$ gives, for every $z \in \mathbb{R}^d$,

$$U_k^\top \mathcal{I}_\beta(z) = \beta U_k^\top U_c \,\mathrm{diag}(\alpha_c) V_c^\top z = 0. \tag{62}$$

Thus $\mathcal{I}_\beta(z)$ is orthogonal to every column of U_k, and hence has no component in the protected output subspace. This proves (iv). □

D IMPLEMENTATION DETAILS

Table 7: Implementation settings for SILO model.

Setting	**Value**
Frozen SSL backbone	XLS-R 300M; $d = 1024$; 24 encoder layers
Adapted projections	query, key, value, output; $\lvert\mathcal{J}\rvert = 96$
Residual / injection ranks	$r = 8$, $r_c = 32$
Preserved energy	$\pi = 0.80$
Projection and gate	P_z: linear $64 \to 1024$ and layer normalization; a_ϕ: MLP $64 \to 64 \to 1$
Amplitude initialization	std. 10^{-4}, signed zero-mean Gaussian
Injection dropout	$p = 0.20$
Factor optimizer	restricted-Stiefel momentum; $\eta = 2 \times 10^{-3}$, $\mu = 0.95$
Other trainable parameters	AdamW; lr 10^{-3}, weight decay 10^{-4}
Scheduler	3-epoch linear warmup, cosine decay to $0.01\times$ base lr
Training	20 epochs; balanced batch 24; gradient clip 1.0
Resource	$1\times$ NVIDIA RTX 5090

D.1 SILO IMPLEMENTATION

This section provides the implementation details for SILO. We follow the notation of Sec. 3 and Appendix B. Tab. 7 summarizes settings for SILO.

Backbone and Adapted Projections. We use `facebook/wav2vec2-xls-r-300m` as the frozen SSL backbone. The encoder contains 24 Transformer layers with hidden width $d = 1024$. SILO uses only 96 adapted projections, with each projection being square, $W_0^{(\ell)} \in \mathbb{R}^{1024\times1024}$. The default residual and injection ranks are $r = 8$ and $r_c = 32$, respectively.

High-Resolution Spectral Encoder and Temporal Alignment. The default E_{spec} uses a high-resolution LFCC front-end. Audio is sampled at 16 kHz. We compute 120 LFCC coefficients from 140 filters using a 1024-point FFT, a 1024-sample analysis window (64 ms), and a 160-sample hop (10 ms). We pretrain E_{spec} on the WaveFake training set as a standalone detector using temporal mean pooling and a linear binary classifier. The classifier is discarded after pretraining, and the spectral-encoder parameters are then loaded into SILO and gradient-frozen. The learned projection $P_z : \mathbb{R}^{64} \to \mathbb{R}^{1024}$ and token-wise layer normalization implement the feature transformation in Eq. 1. Temporal alignment $\mathcal{A}_{T_s\to T}$ is linear interpolation.

Initialization, Protected Subspaces, and Gating. We use $\pi = 0.80$ in the default configuration. The columns of U, V, and U_c are initialized in the orthogonal complements of U_k, V_k, and U_k, respectively. The signed amplitudes are initialized independently as $\alpha_i, \alpha_{c,j} \sim \mathcal{N}(0, 10^{-8})$, i.e., with standard deviation 10^{-4}. The utterance gate $g_\phi(x)$ is computed from the temporally pooled spectral features before P_z. The map a_ϕ is implemented as a two-layer MLP, $\mathbb{R}^{64} \to \mathbb{R}^{64} \to \mathbb{R}$, with a ReLU hidden activation and sigmoid output $g_\phi(x) = \rho(a_\phi(\frac{1}{T_s}\sum_j s_j))$. The output bias is initialized to zero, yielding gate values near 0.5 at initialization. During training, injection dropout is applied independently to each utterance. Specifically, $b \sim \text{Bernoulli}(1-p)$ with $p = 0.20$, and the effective gate is $\beta_b(x) = b\, g_\phi(x)$. The same b and $g_\phi(x)$ are shared across all $\ell \in \mathcal{J}$ for a given utterance. No inverted-dropout rescaling by $1/(1-p)$ is used. At inference, $b = 1$, so $\beta(x) = g_\phi(x)$.

Table 8: **Audio transformation families and settings used in the robustness evaluation**.

Family	Purpose	Transformations and implementation
A: Codec / transmission	Simulates lossy compression and narrowband communication channels.	Real encode–decode round trips: **MP3** at {320,128,64,32,16} kbps; **AAC** at {128,64,32} kbps; **Opus** at {64,32,16} kbps; **G.711** μ-law and A-law encoded at 8 kHz and decoded back to 16 kHz.
B: Noise / room	Tests robustness to additive interference and room propagation effects.	Gaussian white noise at **SNR** {20,10,0} dB; synthetic reverberation with **RT60** {300,600,900} ms; **echoes** with {50,150,300} ms delay using `aecho=0.8:0.9:delay:0.4`.
C: Filtering / EQ / bandwidth	Tests sensitivity to spectral reshaping and bandwidth loss.	**Low-pass** at {3.4,7.0} kHz; **high-pass** at {100,300} Hz; peaking **EQ** at 2 kHz with ±6 dB and $Q = 1$; **down/up-sampling round trips** through {8,4} kHz.
D: Time / pitch	Models active laundering that changes temporal or pitch characteristics.	**Tempo** factors {0.9,1.1} using `atempo`; **pitch shifts** of ±2 semitones using `rubberband` with factors $2^{\pm2/12}$ while preserving duration.
E: Dynamics / amplitude	Tests robustness to level changes and dynamic-range processing.	**Gain** of $+6$ and -12 dB by direct waveform scaling, without clipping or renormalization; **dynamic-range compression** with ratios {4:1,8:1}.
F: Quantization	Tests degradation from coarse amplitude and temporal quantization.	**Linear bit crushing** to {8,4} bits; **sample-and-hold crushing** to effective rates {8,4} kHz using hold factors {2,4}. The latter keeps the waveform at 16 kHz and is distinct from the resampling-based bandwidth reduction in Family C.

D.2 Audio Transformation Techniques

To evaluate robustness to realistic channel degradation and deliberate audio laundering, we apply 42 deterministic transformations grouped into 6 families. All inputs are mono 16-kHz audio, and no waveform normalization is applied. The clean condition is a no-op reference. GSM is excluded from all reported results. Families A–C and E–F preserve linguistic content, whereas Family D changes tempo or pitch and is treated separately as active laundering/evasion.

For reproducibility, all codec and FFmpeg-based transforms operate on float32 mono audio and return a 16-kHz waveform. Noise strength is set from whole-utterance mean power. The reverberation transform uses a fixed synthetic impulse response for a given RT60, and additive-noise realizations are deterministically tied to the input waveform. Hence, the same utterance receives the same transformed signal for every evaluated detector.

Table 9: Full comparison of audio deepfake datasets. ”Variety” reports the number of speakers, generation systems, bonafide (BF) samples, and spoof samples. ”Spoof construction” lists representative generators. ”–” denotes not reported. ”Clinical labels” indicates explicit pathology or clinical-condition annotations, while ”Minutes long”’ denotes datasets whose primary evaluation units are complete multi-minute recordings. MLAAD statistics refer to the v3 configuration used in this work.

Dataset	Bonafide source	Variety	Spoof construction	Duration	Clinical labels	Minutes long
WaveFake (Frank & Schönherr, 2021)	EN/JA; LJSpeech and JSUT read speech	2 speakers; 8 generators; 18,100 bonafides / 117,985 spoofs	MelGAN, MelGAN-L, MB-MelGAN, FB-MelGAN, HiFi-GAN, WaveGlow, PWG, and Conformer+PWG	mean $\approx$ 6.0 s; max N/R	No	No
ASVspoof19-LA (Todisco et al., 2019)	EN; VCTK	107 speakers; 17 distinct TTS/VC systems; 12,483 bonafides / 108,978 spoofs	A01–A19 (17 unique): WaveNet-, WORLD-, Merlin-, and MaryTTS-based TTS; WaveRNN, Griffin–Lim, Vocaine, STRAIGHT, and multiple VC systems	mean/max N/R	No	No
ASVspoof21-DF (Yamagishi et al., 2021)	EN; ASVspoof/VCC-derived speech under media-storage conditions	93 spk.; 100+ generation systems; 22,617 BF / 589,212 spoof	TTS/VC systems drawn from ASVspoof and VCC sources; individual generation systems are largely anonymized; multiple lossy-codec conditions	mean/max N/R	No	No
In-the-Wild (Müller et al., 2022)	EN; public online speech from celebrities and politicians	58 spk.; generator count N/R; 19,963 BF / 11,816 spoof	Found real-world deepfakes collected from public online sources; generator identities not disclosed	mean 4.3 s; max N/R	No	No
MLAAD v3 (Müller et al., 2024)	23 languages; M-AILABS multilingual audiobook speech	spk. N/R; 116 TTS models from 21 architectures; BF count N/R / $\sim$76,000 spoof	XTTS, VITS, YourTTS, Glow-TTS, FastPitch, Tacotron 2, SpeedySpeech, Tortoise, Capacitron, and other Coqui/HuggingFace TTS models	mean $\approx$ 7.8 s; max N/R	No	No
PolyGlotFake (Hou et al., 2024)	7 languages; public online videos; audio-visual setting	spk. N/R; 5 audio-generation pipelines; 766 BF / 14,472 spoof	XTTS; Bark+FreeVC; VALL-E-X; Microsoft TTS+FreeVC; Tacotron+FreeVC	mean 11.79 s; max N/R	No	No
SpeechFake (Huang et al., 2025b)	46 languages; multiple multilingual source corpora	720 spk.; 40 generators; 348,937 BF / 3,338,508 spoof	MelGAN, WaveGlow, PWG, HiFi-GAN, BigVGAN, VITS, StarGAN-VC, Tortoise, StyleTTS2, OpenVoice, GPT-SoVITS, Fish Speech, ChatTTS, CosyVoice, Seed-VC, and commercial APIs	mean/max N/R	No	No
HCFK (Akhtar et al., 2026)	EN/ZH; six healthcare corpora covering depression, Alzheimer's disease, and dysarthria	spk./BF count source-dependent; 7 generators; one paired spoof per BF for each generator	SpeechTokenizer, DAC, EnCodec, SoundStream, FunCodec, AudioDec, and SNAC	mean/max N/R	Yes	No
VDD (ours)	HU; pathological-speech resources	**702 spk.; 7 generators; 706 BF / 4,942 spoof**	**XTTS-v2, F5-TTS-HU, MMS-TTS-HUN, Piper, MKL-VC, Seed-VC, and EZ-VC**	**mean $\approx$ 48.0 s; max 327.7 s**	**Yes**	No
FVD (ours)	HU; ForVoice120 forensic-voice corpus	**120 spk.; 7 generators; 960 BF / 6,720 spoof**	**XTTS-v2, F5-TTS-HU, MMS-TTS-HUN, Piper, MKL-VC, Seed-VC, and EZ-VC**	**mean 5.2 min; max 12.3 min**	No	**Yes**

E Additional Information of VDD and FVD Datasets

E.1 Full Dataset Comparison

Tab. 9 positions VDD and FVD against existing ADD datasets. Existing benchmarks provide broad coverage of generation systems, languages, and recording conditions: WaveFake (Frank & Schönherr, 2021) and ASVspoof (Todisco et al., 2019; Yamagishi et al., 2021) primarily evaluate conventional TTS/VC or neural-vocoder attacks, In-the-Wild (Müller et al., 2022) captures real-world online deepfakes under diverse recording, compression, codec, and post-processing conditions, and MLAAD (Müller et al., 2024), PolyGlotFake (Hou et al., 2024), and SpeechFake (Huang et al., 2025b) substantially expand generator and language diversity.

In contrast, VDD and FVD are designed to complement these evaluation datasets by targeting two underexplored evaluation conditions rather than maximizing dataset scale. VDD targets an important but underexplored failure mode of ADD: pathological speech can differ substantially from the healthy speech that dominates existing training and evaluation corpora. Moreover, synthesis may partially normalize these atypical voice characteristics. As shown in Tab. 10, six of the seven generation systems exhibit higher median harmonics-to-noise ratio (HNR) Boersma et al. (1993) and Cepstral peak prominence (CPPS) Murton et al. (2020) than the bonafide pathological recordings. This indicates that most generated speech exhibits stronger periodic structure and less noise-like phonation than the genuine pathological speech. Synthetic speech may therefore appear acoustically more regular than the genuine speech it imitates. This presents an unusual challenge for detectors trained predominantly on healthy speech: they may associate atypical characteristics of genuine pathological speech with spoofing while treating more regular synthetic speech as closer to their learned distribution of bonafide speech. VDD therefore evaluates whether a detector can separate authenticity from pathology rather than relying on acoustic normality as a shortcut.

Table 10: **Median voice-quality characteristics of VDD bonafide speech and the seven generation systems**. HNR and CPPS are reported in dB, and Δ denotes the difference from the bonafide median. Higher HNR and CPPS indicate stronger periodic structure and less noise-like phonation. These measures characterize acoustic voice quality rather than clinical status.

Attack	System	HNR	ΔHNR	CPPS	ΔCPPS
–	Bonafide	11.26	–	8.07	–
A01	XTTS-v2	11.75	+0.49	9.40	+1.33
A02	F5-TTS-HU	12.31	+1.05	9.34	+1.27
A03	MMS-TTS-HUN	12.87	+1.61	10.03	+1.96
A04	Piper	12.76	+1.50	9.60	+1.53
A05	MKL-VC	10.41	-0.85	7.31	-0.76
A06	Seed-VC	12.03	+0.77	8.81	+0.74
A07	EZ-VC	11.65	+0.39	9.79	+1.72

FVD, in turn, targets long-duration detection. Existing audio deepfake benchmarks predominantly evaluate short utterances or clips, which cannot reveal whether spoofing evidence accumulates across time or how evidence from different parts of a recording should be aggregated. This distinction is important because a detector may be uncertain on an individual short segment while producing a substantially more reliable decision when multiple segments from the same recording are considered. FVD therefore preserves complete recordings ranging from 25 s to 12.3 min and enables systematic evaluation of recording-level aggregation.

E.2 VDD Clinical Composition and Generation Diagnostics

Tab. 11 summarizes the five clinical classes represented by the 706 bonafide recordings. All recordings contain the same read passage, which reduces lexical variation when comparing clinical groups and generation systems. The recordings have a median duration of 46.9 s (mean 49.3 s; maximum 197.6 s). Parkinsonian recordings are longer on average (median 51.5 s) than the healthy-control recordings (median 43.0 s).

Table 11: **Clinical composition of the VDD bonafide recordings.**

Clinical class	Recordings	Female	Male
Healthy control (HC)	228	134	94
Organic dysphonia (OD)	179	98	81
Depression (DE)	131	86	45
Parkinson's disease (PA)	85	40	45
Functional dysphonia (FD)	83	62	21
Total	**706**	**420**	**286**

Intelligibility and speaker consistency. We additionally characterize whether the seven generation systems preserve the intended linguistic content and speaker characteristics. Tab. 12 reports median word error rate (WER), computed against the transcript, and ECAPA speaker-embedding cosine similarity (Desplanques et al., 2020). For the speaker-conditioned TTS systems A01–A02, the similarity to the source speaker is relatively high, particularly for A02. The fixed-voice systems A03–A04, as expected, have much lower similarity to the source speaker. Among the VC

systems, A06 and A07 exhibit substantially higher similarity to their designated target speaker than to the source, whereas A05 retains stronger similarity to the source. A07 also has markedly higher WER than the other systems and is therefore retained as a lower-fidelity but structurally distinct discrete-token VC attack rather than as a high-quality cloning example.

Table 12: **VDD generation diagnostics.** WER is computed using Whisper-large-v3 (Radford et al., 2023) against the ground-truth transcript. ECAPA values are cosine similarities between speaker embeddings. Target similarity applies only to the VC attacks because these systems are conditioned on a separate target identity.

Attack	System	Median WER	ECAPA→source	ECAPA→target
–	Bonafide	–	–	–
A01	XTTS-v2	0.069	0.655	–
A02	F5-TTS-HU	0.118	0.843	–
A03	MMS-TTS-HUN	0.096	0.108	–
A04	Piper	0.069	0.146	–
A05	MKL-VC	0.074	0.644	0.429
A06	Seed-VC	0.115	0.247	0.802
A07	EZ-VC	0.529	0.231	0.791

E.3 FVD Task Structure and Long-Duration Construction

ForVoice120 contains 120 speakers recorded in two sessions and four tasks, forming a complete $120 \times 2 \times 4 = 960$ recording grid. The task structure is important because recording duration and speaking style vary substantially across the corpus. As shown in Tab. 13, Tasks 1 and 2 account for most of the long-duration material, whereas Task 4 offers a particularly useful *content-controlled forensic condition*: every speaker produces the same fixed phrase in both recording sessions. This substantially reduces linguistic-content variation and enables detection to be examined more directly with respect to speaker identity, recording session, and spoof generation.

Table 13: **FVD source-recording statistics by elicitation task.** Each task contains 120 speakers recorded in two sessions.

Task	Recordings	Hours	Median length	Chunks	Speaking style
1	240	39.9	10.0 min	11,073	Free conversation
2	240	30.2	7.3 min	8,576	Collaborative dialogue
3	240	10.8	2.4 min	3,938	Narrative monologue
4	240	2.3	34 s	953	Fixed forensic phrase
All	**960**	**83.2**	**5.4 min**	**24,540**	–

Using the ground-truth time alignment provided by ForVoice120, the 960 recordings are segmented into 24,540 generation chunks. The resulting chunk spans cover 48.34 h of the original timeline, including 15.79 h of short pauses retained within chunks, while another 33.47 h corresponds to gaps between chunks. This distinction is important for reconstruction: TTS systems synthesize the transcribed linguistic content and may not reproduce pauses contained within the corresponding timeline span, whereas VC systems resynthesize the source waveform and therefore preserve its local timing more closely.

Reconstruction fidelity. Because FVD is intended for recording-level rather than isolated chunk-level evaluation, the generated chunks are reassembled according to the original ForVoice120 timeline. For each recording and synthesis system, every generated chunk is first resampled to 16 kHz when necessary, its leading and trailing silence is removed, and its level is matched to that of the corresponding bonafide segment. The chunks are then placed back at their original temporal positions using the boundaries obtained from the time-aligned annotations. Inter-chunk gaps retain their original durations and are filled with background noise synthesized to match the pause characteristics of the corresponding source recording. When a generated chunk is shorter than its assigned timeline span, the remaining interval is padded rather than time-stretching the generated speech. This avoids introducing additional temporal artifacts solely to force alignment. Finally, the reconstructed waveform is matched to the overall loudness of its bonafide counterpart. This procedure preserves the source

Table 14: **FVD generation diagnostics.** Median WER measures linguistic-content preservation using the corpus-provided transcripts as references. ECAPA cosine similarities characterize speaker consistency with respect to the source speech and, for VC systems, the designated target speaker.

Attack	System	Median WER	ECAPA→source	ECAPA→target
–	Bonafide	–	–	–
A01	XTTS-v2	0.000	0.072	–
A02	F5-TTS-HU	0.033	0.080	–
A03	MMS-TTS-HUN	0.026	0.091	–
A04	Piper	0.050	0.037	–
A05	MKL-VC	0.100	0.638	0.385
A06	Seed-VC	0.171	0.100	0.805
A07	EZ-VC	0.571	0.240	0.756

recording's overall duration and temporal structure while retaining the generator-specific timing of the synthesized speech. Consequently, each full-length spoof has the same recording duration and boundaries as its bonafide counterpart, reducing duration as a potential shortcut for detection.

Intelligibility and speaker consistency. We characterize the recording-level generations using the same linguistic and speaker diagnostics as for VDD. Tab. 14 reports the median word error rate (WER), computed with Whisper-large-v3 (Radford et al., 2023) against the ground-truth transcripts, together with ECAPA-TDNN speaker-embedding cosine similarity (Desplanques et al., 2020). The four TTS systems generally preserve the intended linguistic content well: A01–A04 obtain median WERs between 0.000 and 0.050. The speaker-conditioned systems A01–A02 nevertheless show relatively low ECAPA similarity to the source recordings (0.072–0.080), indicating that accurate linguistic reconstruction does not necessarily imply faithful preservation of the source speaker characteristics. Low source similarity is expected for A03 and A04 because these systems use fixed Hungarian synthetic voices rather than speaker-conditioned cloning.

F Additional Experimental Results

F.1 Evaluation on FVD

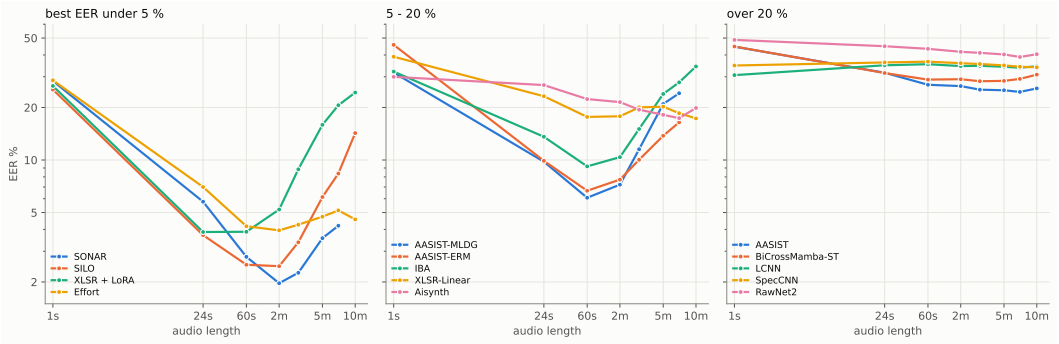


Figure 6: **Recording-level performance on FVD as a function of audio duration (EER%, ↓).** Scores are aggregated over increasing amounts of audio, and methods are grouped by their best achieved EER.

Fig. 6 shows that increasing the amount of available audio can substantially improve deepfake detection, particularly for the stronger detectors. SILO belongs to the best-performing group and reaches its strongest performance after aggregating evidence over approximately 1–2 minutes, demonstrating that spoofing cues that are weak at the segment level can become considerably **more discriminative when accumulated over time**. Similar improvements are observed for several other SSL-based detectors, whereas the weaker systems remain at relatively high EER even with longer recordings. Interestingly, the gains are not monotonic at the longest durations, indicating that simply providing

more audio does not guarantee better detection and that effective temporal evidence aggregation remains an important challenge for long-form ADD.

These results highlight the contribution of FVD as a recording-level benchmark that exposes duration-dependent behavior that cannot be observed from conventional short-clip evaluation. By preserving full-length forensic recordings, FVD enables systematic study of how detectors accumulate and aggregate spoofing evidence over time. This non-monotonic behavior opens an important direction for future work: understanding why additional recording duration does not always translate into better detection.

F.2 Visualization for VDD

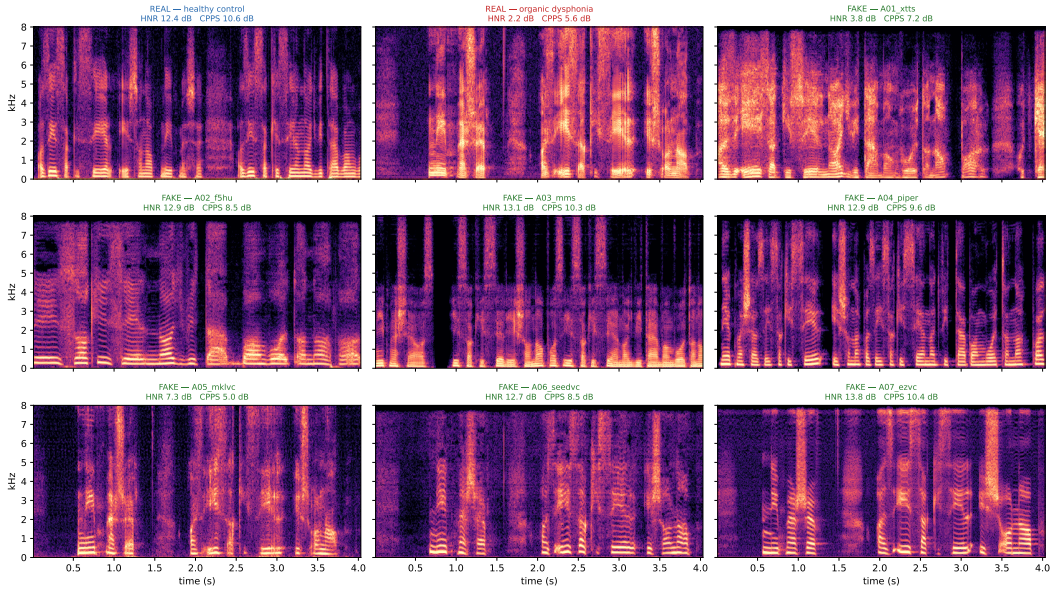


Figure 7: **Representative spectrograms from VDD.** Wideband spectrograms of bonafide healthy and pathological speech together with spoofed samples generated from pathological speech.

Fig. 7 illustrates why pathological speech constitutes a challenging shift for ADD. Compared with the healthy control, the bonafide organic-dysphonia sample exhibits more diffuse spectral energy and less regular harmonic structure, together with substantially lower HNR and CPPS. In contrast, several spoofed samples show clearer harmonic organization and higher voice-quality measures. These examples illustrate that genuine pathological speech can contain irregular spectral and phonatory characteristics that differ markedly from conventional healthy speech, while synthesis may partially reduce such irregularities.

Fig. 8 provides a finer-grained view of this phenomenon. The healthy sample exhibits relatively clear and structured harmonic components, whereas the pathological samples show broader, less regular, and more diffuse spectral patterns. The corresponding mean spectra further highlight differences in the distribution of spectral energy. Such variation within genuine speech can overlap with characteristics that a detector may associate with synthesis artifacts, making pathological speech an important test of whether ADD models distinguish authenticity from pathology-related voice characteristics.

F.3 Robustness to Audio Transformations

We further evaluate robustness under the 42 audio transformations described in Appendix D.2, covering both realistic channel degradation and deliberate audio laundering. Fig. 9 shows that robustness varies substantially across transformation families. SILO remains among the strongest detectors across the three evaluation datasets and is relatively stable under codec/transmission, filtering and bandwidth modification, dynamics, and quantization. In contrast, noise and room effects

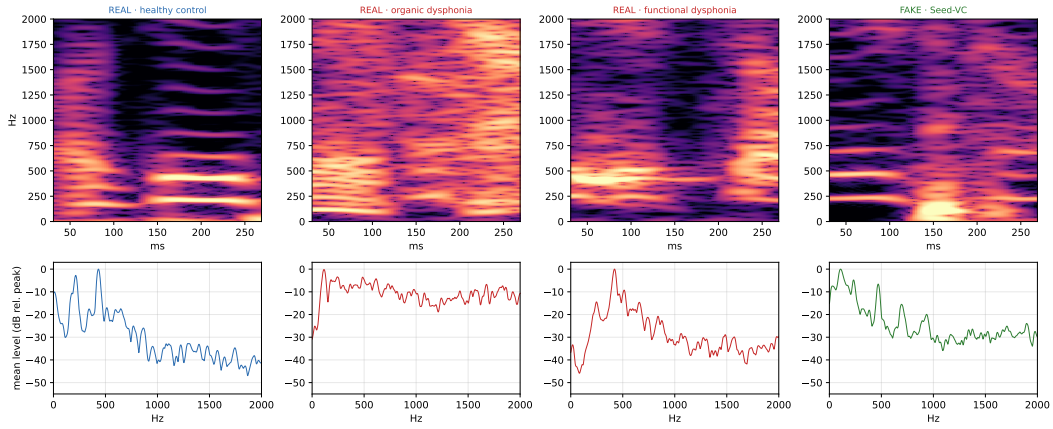


Figure 8: **Fine-grained harmonic structure of representative VDD samples.** Narrowband spectrograms (top) and corresponding mean spectra (bottom) for bonafide healthy, organic-dysphonia, functional-dysphonia, and spoofed speech.

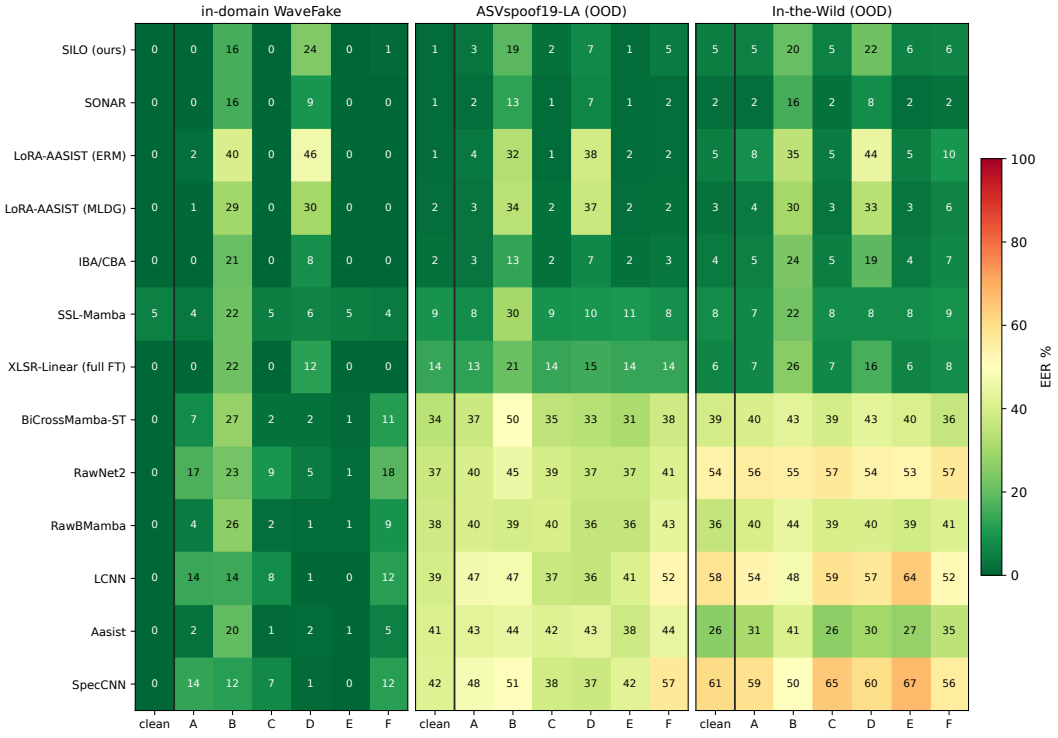
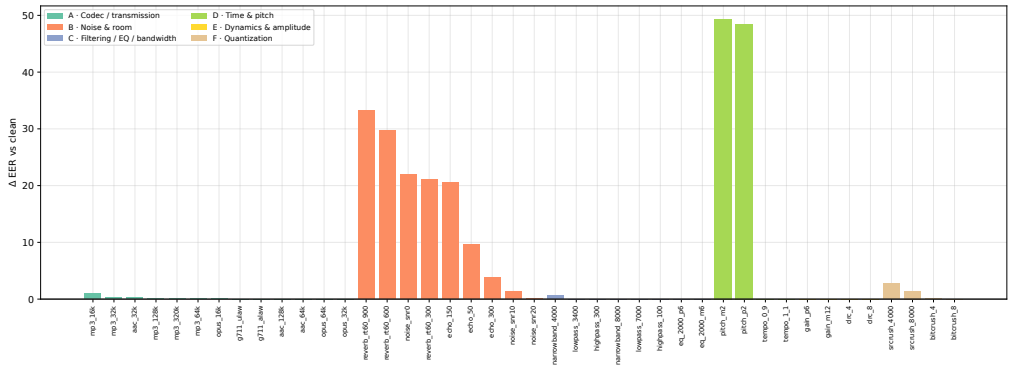


Figure 9: **Robustness to audio transformations on WaveFake, ASVspoof19-LA, and In-the-Wild (EER% ↓).** Columns A–F denote codec/transmission, noise/room, filtering/EQ/bandwidth, time/pitch, dynamics/amplitude, and quantization, respectively. *Clean* denotes untransformed audio. Lower EER indicates greater robustness.

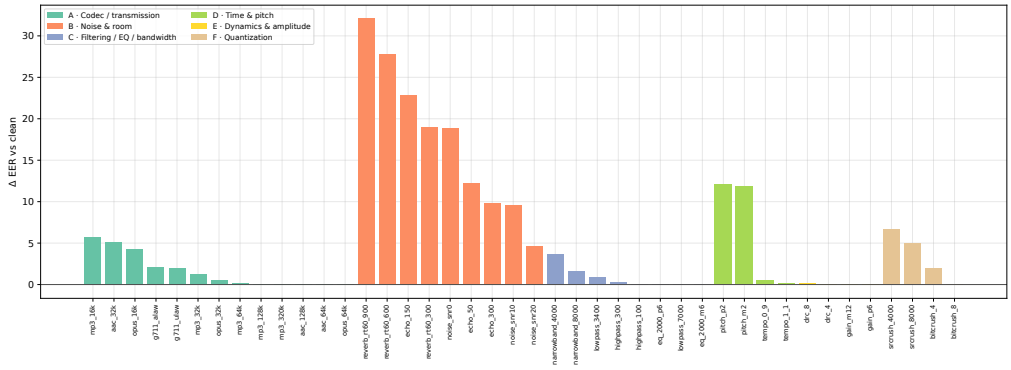
substantially degrade strong detectors such as SILO, SONAR, LoRA-AASIST (ERM/MLDG), and IBA, while time/pitch transformations particularly affect SILO and the two LoRA-AASIST variants.

Fig. 10 provides a finer-grained analysis of SILO. We can observe that the largest degradations are concentrated in severe reverberation, low-SNR noise, and pitch modification. In particular, reverberation with $RT60 = 900$ ms increases EER by approximately $30 - 33\%$ across WaveFake, ASVspoof19-LA, and In-the-Wild, while additive noise at 0 dB SNR increases EER by approximately $16 - 22\%$. These results show that SILO is especially sensitive to transformations that strongly alter spectral and harmonic structure. This observation is consistent with SILO's use of fine-grained

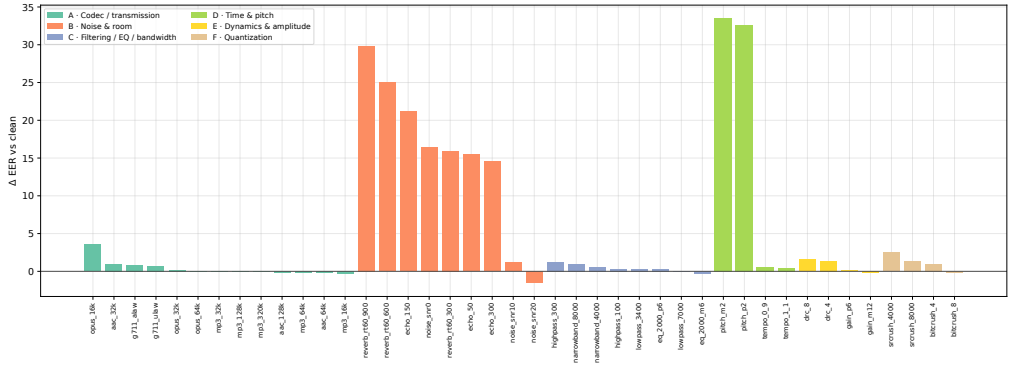
spectral evidence: such evidence improves transfer across unseen generators, but can itself be disrupted by severe acoustic transformations.



(a) WaveFake (in-domain).



(b) ASVSpoof19-LA (OOD).



(c) In-the-Wild (OOD).

Figure 10: **Sensitivity of SILO to individual audio transformations.** Change in EER relative to clean audio (ΔEER, lower is better) is shown for (a) WaveFake, (b) ASVspoof19-LA, and (c) In-the-Wild. Transformations are grouped into the six families A–F defined in Appendix D.2.

F.4 More Generalization Results on Unseen Generators

Fig. 11 complements the evaluation on modern generators in Sec. 5.1.2 by considering a diverse set of classical synthesis architectures, including VITS, Tacotron2, OverFlow, SpeechT5, FastPitch, Capacitron, Glow-TTS, and SpeedySpeech. SILO maintains consistently low, single-digit EER across all evaluated generators, with particularly large improvements over SSL-based baselines such as XLS-R + Linear, XLS-R + LoRA, Effort, and SONAR on challenging systems including MMS-TTS, Tacotron2, OverFlow, SpeechT5, and FastPitch. Classical spectral detectors remain competitive on

several generators, further supporting the transferability of spectral artifacts across synthesis families. Importantly, this figure emphasizes the role of spectral injection in SILO, as SILO substantially closes this gap. Together with the modern-generator results in Fig. 4, these findings show that SILO's generalization is not tied to a particular generation paradigm or generation era, but extends across both classical and recent synthesis systems.

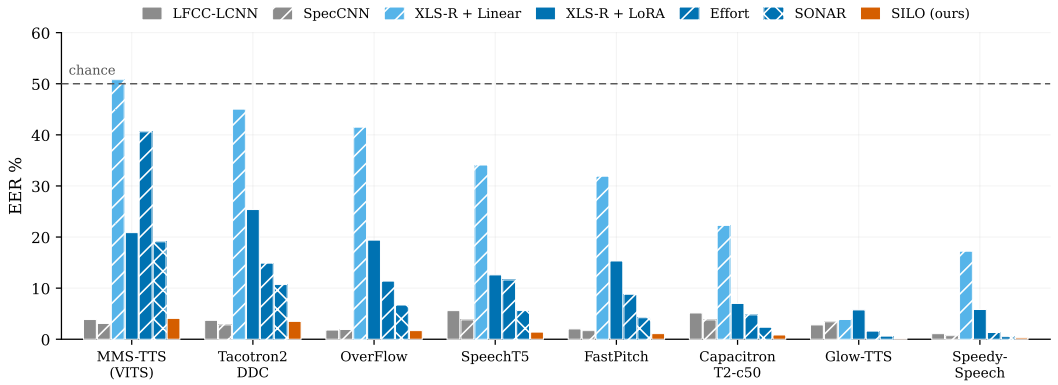


Figure 11: **Generalization on unseen classical generators on MLAAD (EER% ↓).** SILO maintains low EER across diverse classical TTS architectures and substantially improves over SSL-based baselines.

F.5 Performance on Unseen Language

We further investigate whether cross-dataset degradation can be attributed to language differences rather than only to unseen generators. To better separate these factors, we intentionally evaluate on CVoiceFake Li et al. (2024a) using Multi-band MelGAN-v2, Style MelGAN-v1, and Parallel WaveGAN-v1, whose generation mechanisms closely match those represented in the WaveFake training set. Consequently, English serves as a reference condition, while German, French, Italian, and Chinese introduce a language shift primarily. Fig. 12a shows that changing the language does not consistently degrade detection performance. Several detectors achieve lower EER on German (DE) and Chinese (ZH-CN) than on English (EN), despite EN being the language represented in the WaveFake training data. This suggests that **language shift alone is not necessarily detrimental to ADD**, and that detection difficulty depends on the interaction between language-specific speech characteristics and the cues exploited by each detector.

Figs. 12b-12c extend the analysis to more challenging multilingual settings, PolyGlotFake Hou et al. (2024) and MLAAD Müller et al. (2024), in which language shift is accompanied by larger changes in generators and data domains. On PolyGlotFake, achieves the lowest or near-lowest EER across all evaluated languages and, importantly, its performance varies only modestly from EN to FR, ES, ZH, and RU. In contrast, several baselines exhibit much larger degradation, with EER approaching or exceeding 50% for some languages (e.g., BiCrossMamba-ST, AIsynth_Voice, IBA, RawBMamba). The same behaviour can also be observed on MLAAD, as some detectors increase the EER to near 50%, while SILO still achieves the lowest EER.

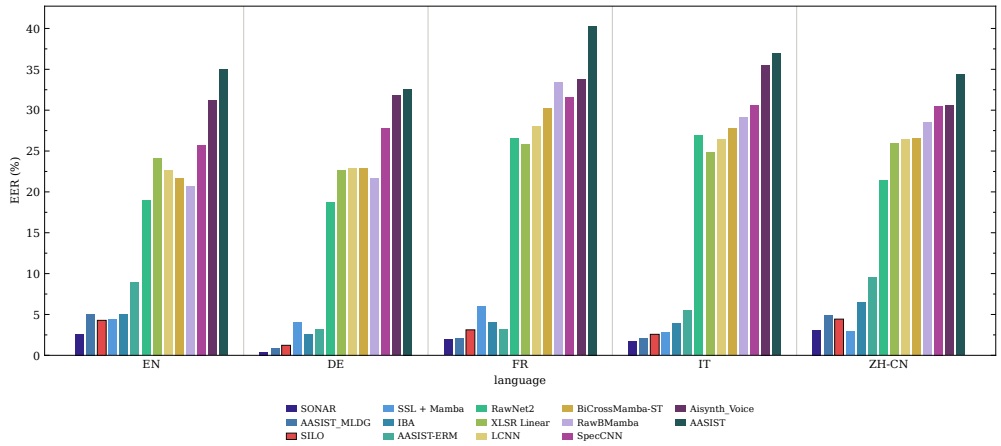
These experimental results reveal that language shift alone is not necessarily detrimental, as shown by CVoiceFake, while multilingual evaluation becomes substantially more difficult when language variation is coupled with changes in generators and data domains.

F.6 Additional Analysis

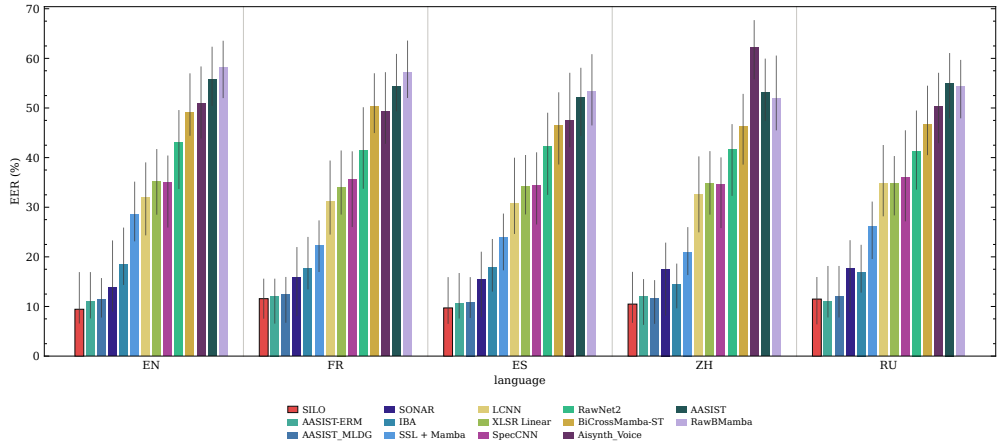
Compatibility with other classical spectral front-ends. We evaluate on 3 other spectral families: Log Power Spectrum (LPS) (Cohen et al., 2022), Constant-Q Cepstral Coefficients (CQCC) (Todisco et al., 2016), and Mel-Frequency Cepstral Coefficients(MFCC) (Davis

Table 15: **Compatibility with classical spectral front-ends (EER%, ↓).**

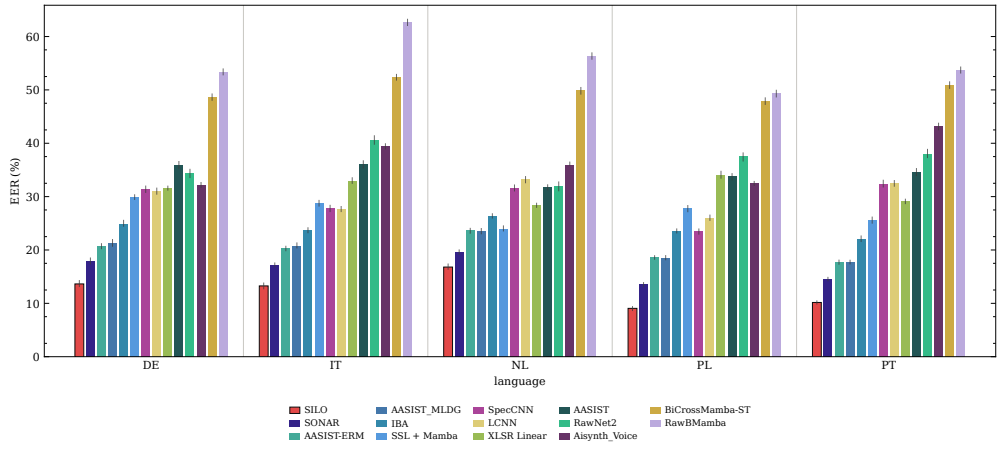
Family	MLAAD	ITW	ASVS-19	ASVS-21
No injection	24.67	**2.80**	1.55	2.15
LFCC	**14.38**	4.35	**1.02**	**1.56**
LPS	15.56	5.26	2.01	2.02
CQCC	17.09	5.73	2.26	2.86
MFCC	18.09	5.20	1.43	2.27



(a) CVoiceFake.



(b) PolyGlotFake.



(c) MLAAD.

Figure 12: **Performance under unseen-language shifts (EER% ↓).** All detectors are trained on WaveFake and evaluated language-wise on multilingual deepfake datasets.

& Mermelstein, 1980). Tab. 15 shows that SILO is not specific to LFCC (Sahidullah et al., 2015): all tested spectral representations substantially

improve MLAAD generalization over using no spectral injection. However, their behavior differs across domains, indicating that the choice of spectral evidence affects the generalization trade-off. LFCC provides the strongest overall balance, with the best results on MLAAD and both ASVspoof benchmarks, and is therefore used as our default spectral front-end.

Learned gate vs. fixed scalar gate. To examine the effect of utterance-adaptive gating, we replace the learned gate $g_\phi(x)$ with a fixed scalar $\beta \in \{0, 0.25, 0.5, 0.75, 1\}$, shared by all utterances. Here, $\beta = 0$ completely disables spectral injection, whereas $\beta = 1$ always applies the full spectral-injection branch. The learned setting uses the input-dependent gate described in Sec. 3.3.

Table 16: **Ablation of the spectral gate.** Comparison between fixed scalar injection strengths β and the learned utterance-adaptive gate. EER (%) ↓ and AUC (%) ↑.

	ASVS-19		ITW		MLAAD	
β	EER ↓	AUC ↑	EER ↓	AUC ↑	EER ↓	AUC ↑
0	6.23	98.19	**1.43**	**98.42**	25.25	79.79
0.25	3.73	99.10	2.65	97.32	20.69	84.36
0.50	2.24	99.49	3.41	96.94	19.78	84.57
0.75	1.22	99.49	6.11	92.77	20.07	83.22
1	1.21	99.88	7.05	90.59	15.89	86.92
Learned	**1.02**	**99.90**	4.35	97.50	**14.38**	**89.50**

From Tab. 16, we can observe that increasing the fixed injection strength generally improves performance over $\beta = 0$, confirming the benefit of spectral evidence. However, the preferred scalar varies across datasets, indicating that a single fixed injection strength cannot provide the best cross-domain trade-off. In contrast, the learned gate achieves the lowest EER on all three evaluated datasets, reaching 1.02% on ASVS-19, 4.35% on ITW, and 14.38% on MLAAD. These results support the use of utterance-adaptive gating: rather than applying the same amount of spectral evidence to every input, SILO learns to adjust its contribution according to the current utterance.

G LIMITATIONS

Our study focuses on cross-dataset generalization under a common WaveFake training setting, and broader multi-source training scenarios remain for future evaluation. VDD and FVD intentionally provide controlled evaluations of pathological speech and long-form recordings, but currently cover Hungarian speech and a fixed set of seven generation systems. Extending these datasets to additional languages and generators would further broaden their scope. Finally, FVD is designed to expose duration-dependent behavior rather than introduce a dedicated long-context aggregation mechanism, which we leave for future work.