

A Training-Free Projection-Guided Conditional Flow-Matching System for VoicePrivacy 2026

Xiang Shi¹, Han Zhu², Ming Li³, and Xiaoxiao Miao^{*1}

¹Duke Kunshan University

²Xiaomi Corporation

³School of Artificial Intelligence, The Chinese University of Hong Kong, Shenzhen

¹{xiang.shi, xiaoxiao.miao}@dukekunshan.edu.cn, zhuhan3@xiaomi.com, ming.li.cuhksz@gmail.com

Abstract

Voice anonymization aims to conceal speaker identity while preserving linguistic and paralinguistic information, yet many existing methods require dedicated training. We propose a training-free, two-stage anonymization method for pretrained flow-matching voice conversion systems. In the condition-construction stage, the source content representation is regenerated under a randomly selected speaker context to reduce residual source-speaker information while preserving linguistic content. In the speech-generation stage, the source speech is used to identify the generation direction associated with the original speaker, and the source-aligned component is suppressed while useful content guidance is retained. The proposed method modifies only inference-time conditioning and guidance, without updating any pretrained parameters. Experiments on VoicePrivacy 2026 Track 1 and Track 2 show that the two-stage method improves speaker privacy while maintaining competitive word error rate and emotion recognition performance.

Index Terms: voice anonymization, speaker privacy, voice conversion, flow matching, classifier-free guidance

1. Introduction

Speech contains not only linguistic content but also private information such as speaker identity, emotion, health, and social background [1]. Voice anonymization aims to prevent reliable source-speaker re-identification while preserving linguistic and paralinguistic information required by downstream applications [2]. The VoicePrivacy Challenge provides common datasets, baselines, and evaluation protocols for this task [3, 4]. Recent editions have introduced stronger attacker models and additional utility measures, including emotion preservation [5–9].

Existing anonymization systems broadly follow signal-processing or neural synthesis approaches. Signal-processing methods modify acoustic properties such as formants or speaking rate, but may distort speech and remain vulnerable to strong attackers [10–13]. Neural methods reconstruct speech from text, disentangled representations, transformed speaker embeddings, or discrete speech tokens [2, 14–19]. More recently, diffusion and flow-matching models have enabled high-quality and controllable anonymization [20–22]. However, most such systems require anonymization-specific training or additional learned components.

We propose a training-free anonymization system for pretrained conditional flow-matching voice conversion models. The method modifies two parts of the inference process. First, the source-content condition is regenerated under an independently selected speaker context to reduce its dependence on the source speaker. Second, the source utterance is used to identify a source-related generation direction, and the aligned component of token guidance is suppressed while useful linguistic guidance is retained. No pretrained parameters are updated.

We implement the system on Seed-VC v2 [23]. As shown in Figure 1, random-target conditioning is applied during condition construction, while projection-guided classifier-free guidance is applied during conditional flow-matching generation. Experiments under the VoicePrivacy 2026 Track 1 English protocol and Track 2 multilingual protocol show improved speaker privacy while maintaining competitive word error rate and emotion recognition performance.

2. Preliminaries

2.1. Seed-VC v2

Seed-VC v2 combines autoregressive (AR) speech-token conversion with conditional flow-matching speech generation.

¹ In standard voice conversion, source narrow-band tokens provide the linguistic sequence, while target narrow- and wide-band tokens provide autoregressive context and prefix. The resulting conditions are

$$\begin{aligned}\hat{\mathbf{z}}_w &= G_{AR}(\mathbf{z}_n^{\text{src}}, \mathbf{z}_n^{\text{tar}}, \mathbf{z}_w^{\text{tar}}), \\ \boldsymbol{\mu}_{AR} &= H(\hat{\mathbf{z}}_w), \quad \boldsymbol{\mu}_{\text{prm}} = H(\mathbf{z}_w^{\text{tar}}), \\ \boldsymbol{\mu} &= \text{Concat}_t(\boldsymbol{\mu}_{\text{prm}}, \boldsymbol{\mu}_{AR}).\end{aligned}\tag{1}$$

Here, $H(\cdot)$ regulates token sequences to their corresponding Mel-frame lengths. The target utterance also provides an acoustic prompt $\mathbf{p} = M(x^{\text{tar}})$ and a global style embedding $\mathbf{s} = E_s(x^{\text{tar}})$. The pair $(\boldsymbol{\mu}_{\text{prm}}, \mathbf{p})$ describes an aligned reference segment, $\mathbf{s}$ represents global target timbre, and $\boldsymbol{\mu}_{AR}$ specifies the source content to generate.

At each flow-matching step, the shared velocity estimator is evaluated under three condition sets:

$$\begin{aligned}\mathbf{v}_f &= \mathbf{v}_\theta(\mathbf{x}_t, t; \mathbf{p}, \mathbf{s}, \boldsymbol{\mu}), \\ \mathbf{v}_c &= \mathbf{v}_\theta(\mathbf{x}_t, t; \mathbf{0}, \mathbf{0}, \boldsymbol{\mu}), \\ \mathbf{v}_u &= \mathbf{v}_\theta(\mathbf{x}_t, t; \mathbf{0}, \mathbf{0}, \mathbf{0}).\end{aligned}\tag{2}$$

^{*}Corresponding author.

¹Seed-VC v2 checkpoints: <https://huggingface.co/Plachta/Seed-VC/tree/main/v2>.

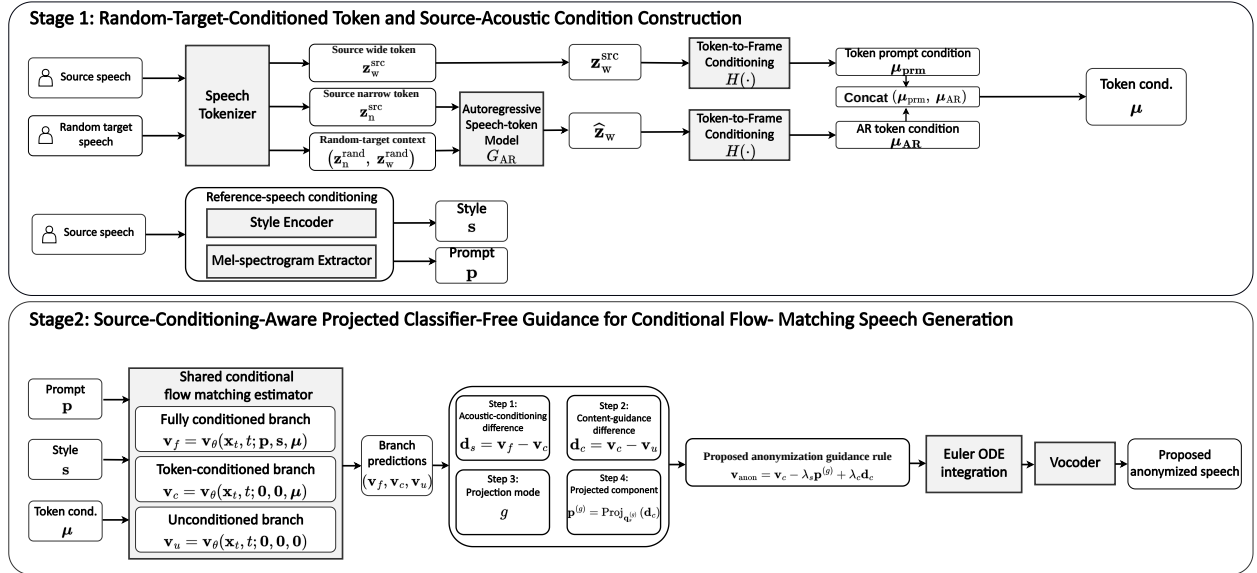


Figure 1: Overview of the proposed training-free anonymization system. Random-target context is used to construct the source-content token condition in Stage 1. In Stage 2, projected classifier-free guidance (CFG) suppresses the component of token guidance aligned with source-acoustic conditioning before Euler integration and vocoding.

These are the fully conditioned, token-conditioned, and unconditional predictions, respectively. Standard voice conversion combines them through classifier-free guidance and then generates a Mel-spectrogram using Euler integration; a pretrained BigVGAN v2 vocoder converts it to waveform audio [24].²

2.2. Built-in Anonymization

The released Seed-VC v2 system also includes a training-free anonymization mode. It removes the target autoregressive context and disables acoustic-prompt and style conditioning:

$$\begin{aligned} \hat{\mathbf{z}}_w^{\text{builtin}} &= G_{\text{AR}}(\mathbf{z}_n^{\text{src}}; \emptyset, \emptyset), \\ \boldsymbol{\mu}_{\text{AR}}^{\text{builtin}} &= H(\hat{\mathbf{z}}_w^{\text{builtin}}), \\ \boldsymbol{\mu}_{\text{prm}}^{\text{builtin}} &= H(\mathbf{z}_w^{\text{src}}), \quad \mathbf{p} = \mathbf{0}, \quad \mathbf{s} = \mathbf{0}. \end{aligned} \quad (3)$$

The complete token condition is $\boldsymbol{\mu}^{\text{builtin}} = \text{Concat}_t(\boldsymbol{\mu}_{\text{prm}}^{\text{builtin}}, \boldsymbol{\mu}_{\text{AR}}^{\text{builtin}})$. Generation uses

$$\mathbf{v}^{\text{builtin}} = \mathbf{v}_c^{\text{builtin}} + \gamma_{\text{joint}}(\mathbf{v}_c^{\text{builtin}} - \mathbf{v}_u), \quad (4)$$

where $\mathbf{v}_c^{\text{builtin}} = \mathbf{v}_\theta(\mathbf{x}_t, t; \mathbf{0}, \mathbf{0}, \boldsymbol{\mu}^{\text{builtin}})$. This mode serves as the training-free baseline for our system.

3. Proposed Method

The built-in Seed-VC anonymization mode already provides a training-free way to anonymize speech without updating the pretrained model. However, it still leaves room for improvement in both stages. In Stage 1, the target-related AR inputs are simply removed, so the model does not exploit its ability to regenerate the source content under an alternative speaker context. As a result, the token condition is still largely determined by the source input and may carry residual source-speaker information entangled with the content. In Stage 2, the guidance rule $\mathbf{v}_c^{\text{builtin}} + \gamma_{\text{joint}}(\mathbf{v}_c^{\text{builtin}} - \mathbf{v}_u)$ reduces the direct use of source acoustic conditions, but it neither identifies and removes the source-related information carried in the token guidance nor steers the generation trajectory away from the source

speaker. These observations motivate our two modifications: random-target conditioning in Stage 1, which regenerates the source content under an independently sampled speaker context to weaken residual source-speaker information, and source-conditioning-aware projected CFG in Stage 2, which actively suppresses source-aligned directions during generation. Figure 1 provides an overview of the proposed two-stage anonymization system, including random-target-conditioned token construction in Stage 1 and source-conditioning-aware projected CFG in Stage 2.

3.1. Stage 1: Random-Target-Conditioned Token and Source-Acoustic Condition Construction

In Stage 1, we change only the AR-generated part of the token condition. As shown in Stage 1 of Figure 1, the complete token condition contains two parts: the source-derived token-prompt condition $\boldsymbol{\mu}_{\text{prm}}$ and the random-target-conditioned AR token condition $\boldsymbol{\mu}_{\text{AR}}$. The former describes the prompt region, whereas the latter describes the source content to be generated.

To reduce source-speaker information in the generated source-content tokens, we sample an independent random-target utterance and use its narrow- and wide-band tokens as the AR context and prefix. The source narrow-band tokens still provide the linguistic sequence to be converted. The shared AR model therefore predicts the source-portion wide-band tokens as

$$\hat{\mathbf{z}}_w = G_{\text{AR}}(\mathbf{z}_n^{\text{src}}; \mathbf{z}_n^{\text{rand}}, \mathbf{z}_w^{\text{rand}}). \quad (5)$$

The generated sequence is then converted to the Mel-frame rate as $\boldsymbol{\mu}_{\text{AR}} = H(\hat{\mathbf{z}}_w)$. Thus, $\boldsymbol{\mu}_{\text{AR}}$ still carries the source linguistic content, but its token realization is guided by an unrelated speaker context.

The token-prompt condition is constructed separately from the source wide-band tokens as $\boldsymbol{\mu}_{\text{prm}} = H(\mathbf{z}_w^{\text{src}})$. We deliberately keep $\boldsymbol{\mu}_{\text{prm}}$ source-derived because the same source utterance also provides the acoustic prompt $\mathbf{p}$. This keeps the token description of the prompt region aligned with its Mel-spectrogram. If the token-prompt condition were instead taken from the random target while the acoustic prompt remained source-derived, the model would receive random-target prompt tokens together with a source acoustic example.

²BigVGAN v2 checkpoint: https://huggingface.co/nvidia/bigvgan_v2_22khz_80band_256x.

This prompt mismatch differs from the aligned conditioning used during Seed-VC training. In our preliminary tests, such a mismatch sometimes improved speaker separation, but it also made generation unstable and substantially degraded audio quality and increased WER. We therefore use the source utterance for both the token-prompt condition and the acoustic prompt.

The complete token condition is formed by temporal concatenation as $\mu = \text{Concat}_t(\mu_{\text{prm}}, \mu_{\text{AR}})$. This operation is shown explicitly by the *Concat* block in Figure 1. The two parts are first regulated to their respective Mel lengths and then concatenated along the temporal dimension.

In parallel, the source utterance is passed through the same Mel-spectrogram extractor and CAM++ style encoder used by Seed-VC, giving $\mathbf{p} = M(x^{\text{src}})$ and $\mathbf{s} = E_s(x^{\text{src}})$, where $M(\cdot)$ denotes prompt-Mel extraction and $E_s(\cdot)$ denotes the style encoder. The source-derived μ_{prm} and $\mathbf{p}$ form an aligned prompt example, while $\mathbf{s}$ describes the overall source timbre. However, these source conditions are not used as desired output attributes. They are retained only to expose the effect of source-acoustic conditioning in Stage 2.

Because the complete token condition retains a source prompt part and AR generation starts from source tokens, Stage 1 reduces but cannot guarantee the removal of all source-speaker information.

3.2. Stage 2: Source-Conditioning-Aware Projected Classifier-Free Guidance for Conditional Flow-Matching Speech Generation

In Stage 2, illustrated in the lower part of Figure 1, we use the source prompt and style embedding only to identify the generation direction associated with the source speaker. At each ODE step, the difference between the fully conditioned and token-conditioned predictions shows how the source acoustic conditions change the current generation state. The difference between the token-conditioned and unconditional predictions shows the effect of the complete token condition μ , which contains both μ_{prm} and μ_{AR} . We project this token-guidance direction onto the source-conditioning direction and suppress the aligned component.

Accordingly, the source prompt $\mathbf{p}$ and style embedding $\mathbf{s}$ serve only as a reference. They identify what should be suppressed rather than what should be reconstructed. The remaining token guidance continues to support linguistic generation, so sampling starts from the token-conditioned field $\mathbf{v}_c$, not from the fully conditioned field $\mathbf{v}_f$.

Concretely, as shown by the three branches in Figure 1, at every ODE step, the shared estimator is evaluated under the same three condition sets as in Eq. (2), using the $(\mu, \mathbf{p}, \mathbf{s})$ constructed in Stage 1 to obtain $\mathbf{v}_f$, $\mathbf{v}_c$, and $\mathbf{v}_u$. We then form

$$\mathbf{d}_s = \mathbf{v}_f - \mathbf{v}_c, \quad \mathbf{d}_c = \mathbf{v}_c - \mathbf{v}_u, \quad (6)$$

where $\mathbf{d}_s$ is the source-conditioning direction contributed by the source acoustic prompt and style embedding, and $\mathbf{d}_c$ is the guidance direction contributed by the complete token condition. Their alignment identifies the part of token guidance that changes the acoustic state in a way consistent with reintroducing source conditioning. We do not assume that $\mathbf{d}_s$ is a pure speaker direction or that $\mathbf{d}_c$ is a pure content direction.

The four projection operations shown in Figure 1 consist of constructing the source-conditioning difference, constructing

the token-guidance difference, selecting the temporal projection mode, and computing the projected component. Both $\mathbf{d}_s$ and $\mathbf{d}_c$ have shape $C \times T$, where C is the feature dimension and T is the number of frames. The projection reference $\mathbf{q}_{s,t}^{(g)}$ is built from $\mathbf{d}_s$ at one of three temporal granularities,

$$\mathbf{q}_{s,t}^{(g)} = \begin{cases} \mathbf{d}_{s,t}, & g = \text{token}, \\ \frac{1}{T} \sum_{\tau=1}^T \mathbf{d}_{s,\tau}, & g = \text{global}, \\ \frac{1}{K} \sum_{\tau \in \mathcal{W}_t} \mathbf{d}_{s,\tau}, & g = \text{smooth}, \end{cases} \quad (7)$$

where *Token* uses the source-conditioning vector at the same frame, *Global* uses one utterance-level mean, and *Smooth* uses a local sliding-window mean over $\mathcal{W}_t = \{t-10, \dots, t+10\}$.³ These choices respectively preserve frame-level variation, remove local variation, or provide an intermediate smoothed reference.

The component of token guidance aligned with the selected reference is

$$\mathbf{p}_t^{(g)} = \frac{\langle \mathbf{d}_{c,t}, \mathbf{q}_{s,t}^{(g)} \rangle}{\max\left(\left\|\mathbf{q}_{s,t}^{(g)}\right\|_2^2, \epsilon\right)} \mathbf{q}_{s,t}^{(g)}. \quad (8)$$

The vector being projected is always $\mathbf{d}_{c,t}$; only the temporal construction of the reference changes across modes. We treat $\mathbf{p}_t^{(g)}$ as a proxy for source-speaker information remaining in the token pathway.

The proposed guidance rule removes this component and then reinforces the token guidance,

$$\mathbf{v}_{\text{anon}} = \mathbf{v}_c - \lambda_s \mathbf{p}_t^{(g)} + \lambda_c \mathbf{d}_c, \quad (9)$$

where λ_s controls source-aligned suppression and λ_c controls token guidance. Using $\mathbf{v}_c$ as the base field prevents the source prompt and style from directly reconstructing the source voice, while the difference between $\mathbf{v}_f$ and $\mathbf{v}_c$ still indicates which direction should be removed. The Euler integration and vocoder stages are shown at the end of Figure 1. Specifically, the shared Euler solver integrates $\mathbf{x}_{t_{k+1}} = \mathbf{x}_{t_k} + \Delta t_k \mathbf{v}_{\text{anon}}(\mathbf{x}_{t_k}, t_k)$, after which the pretrained vocoder renders the final waveform.

4. Experiments

Our experiments proceed in three steps. First, we construct a small screening subset containing 30 utterances selected from the VPC LibriSpeech development-clean set [25], denoted *Libri_dev*. The screening subset contains 30 speakers, with one utterance selected per speaker. For each speaker, we choose the longest clearly spoken utterance to reduce measurement noise in WER and source-speaker dissimilarity during the dense grid search.⁴ On this subset, we search for the best (λ_s, λ_c) pair for each combination of random-target condition construction and projection mode. Second, with the selected scale pairs fixed, we evaluate all systems on the full VPC

³ $K = 21$: the current frame plus the ten preceding and ten following frames, with reflection padding at utterance boundaries.

⁴The subset is used only for quick screening and was empirically sufficient, as the selected settings generalized well to the full development and evaluation sets. Using the full development set could potentially yield better settings, but would make the dense (λ_s, λ_c) sweep substantially more expensive.

Table 1: Track 1 development-set ablation results under the lazy-informed attack. RT means that μ_{AR} uses random-target context. Higher Lazy-EER and UAR are better, while lower WER is better. The highest Lazy-EER is shaded.

System	RT	λ_s	λ_c	Lazy-EER	WER	UAR
Built-in		0.7	0.7	33.15	4.18	41.85
Global		0.9	2.8	33.90	4.20	39.21
Smooth-21		1.0	3.0	34.20	4.39	40.35
Token		0.2	2.4	33.48	4.22	40.30
Global	✓	1.2	2.9	34.61	4.42	40.13
Smooth-21	✓	1.1	2.8	35.09	4.53	40.38
Token	✓	0.1	1.3	34.72	4.36	39.95

2026 Track 1 development and evaluation sets. These full-set experiments serve as ablation studies for the random-target-conditioned construction of μ_{AR} introduced in Section 3.1 and the projected guidance proposed in Section 3.2. Third, to evaluate multilingual generalization, we test two fixed systems on the Track 2 development and evaluation sets: the Seed-VC built-in anonymization mode and one configuration of our proposed method that achieved strong performance on both the Track 1 development and evaluation sets. Their settings are kept unchanged, with no additional tuning on Track 2.

4.1. Evaluation Protocol, Datasets, and Settings

We follow the VoicePrivacy Challenge 2026 Track 1 (English) and Track 2 (multilingual) evaluation protocol [9] and use the challenge-provided datasets, attack models, and automatic speech recognition (ASR) and speech emotion recognition (SER) evaluation models without modification. Speaker privacy is measured by equal error rate (EER) under lazy-informed and semi-informed attacks, where a higher value indicates stronger protection against speaker re-identification. We denote the same EER metric under these two attack settings as Lazy-EER and Semi-EER, respectively. Linguistic-content preservation is measured by word error rate (WER, lower is better), and emotion preservation is measured by unweighted average recall (UAR, higher is better).

The random-target pool introduced in Section 3 contains one short utterance (3–6 s) from each of 500 speakers in the LibriTTS train-other-500 subset [26]. For every source utterance used by a random-target-conditioned system, one target utterance is sampled from this pool with replacement. The sampled target utterance provides the narrow-band context and wide-band prefix used to construct the AR-generated part μ_{AR} . The token-prompt part μ_{prm} remains source-derived so that it stays aligned with the source acoustic prompt. A fixed random seed is used, and the same source–target assignment is retained across all projection modes and scale settings to ensure a controlled comparison.

4.2. Results

Guidance Scale Grid Search on Libri_{dev}. We take the subset Libri_{dev} to select the operating point of each system. For each of the three projection modes (Global, Smooth-21, and Token), we test both target-free AR construction and random-target-conditioned AR construction. In both cases, the token-prompt condition μ_{prm} is extracted from the source utterance. This gives six configurations in total.

For every configuration, both λ_s and λ_c are swept from 0 to 3 with a step size of 0.1. The source utterances and, when applicable, their random-target assignments remain fixed throughout the search. We jointly inspect the anonymization score and WER and select the pair that provides the best

privacy–utility balance for that configuration. The selected values are reported in Table 1. After selection, all scale pairs are fixed and no further tuning is performed on the complete Track 1 development or evaluation sets.

Table 1 reports the Track 1 development-set ablation results, while Table 2 compares the selected Smooth-21 system with the official Track 1 baselines and built-in anonymization.

Effect of Random-Target Conditioning. Table 1 isolates the effect of the Stage 1 random-target-conditioned construction of μ_{AR} introduced in Section 3.1. For each projection mode, the configuration using random-target AR context is compared with its counterpart using target-free AR generation. The source-derived μ_{prm} is unchanged in each pair. Random-target conditioning improves development Lazy-EER from 33.90% to 34.61% for Global, from 34.20% to 35.09% for Smooth-21, and from 33.48% to 34.72% for Token. WER and UAR remain within a similar range.

The consistent improvement across all three projection modes shows that the benefit does not depend on one particular temporal granularity. Guiding AR generation with an independently sampled target utterance reduces the source dependence of the AR-generated part μ_{AR} . We therefore retain random-target-conditioned AR construction in the selected system used for the subsequent Track 1 semi-informed comparison.

Effect of the Projection Method. Table 1 also evaluates the projected speech-generation rule introduced in Section 3.2. The three configurations without random-target context use the same target-free construction of μ_{AR} and the same source-derived μ_{prm} as the built-in Seed-VC anonymization mode. Their main difference is therefore in Stage 2. The built-in mode uses the complete token condition with prompt and style set to zero, whereas the proposed systems additionally use source prompt and style to estimate the source-conditioning direction and then remove the aligned part of token guidance.

All three projection variants improve development Lazy-EER over the built-in mode: Global, Smooth-21, and Token reach 33.90%, 34.20%, and 33.48%, respectively, compared with 33.15% for built-in anonymization. Their WER remains close to that of the built-in system. This comparison shows that the privacy gain is not produced only by random-target conditioning; actively suppressing the source-aligned component during speech generation is itself beneficial. Smooth-21 gives the strongest privacy result among the three projection granularities.

Overall Comparison with Track 1 Baselines. Table 2 compares the selected random-target-conditioned Smooth-21 system with the official Track 1 baselines and the built-in Seed-VC anonymization mode. On the evaluation set, Smooth-21 reaches a Lazy-EER of 29.39% and a Semi-EER of 21.51%, both higher than the corresponding results of the built-in anonymization mode. It also exceeds the strongest official baseline B5 in both evaluation EER measures, while maintaining a comparable WER. Smooth-21 also achieves the highest development EERs, confirming its effectiveness against a semi-informed attacker while preserving competitive utility.

Overall Comparison with Track 2 Baselines. Table 3 compares the official Track 2 baselines and the Seed-VC built-in anonymization mode with the selected Global random-target-conditioned configuration, which achieved strong performance on both the Track 1 development and evaluation sets. For

Table 2: Track 1 comparison with the official baselines under lazy-informed and semi-informed attacks. Privacy is measured by EER (%), and utility by WER (%) and UAR (%). A checkmark indicates that μ_{AR} uses random-target context. A slash denotes a non-applicable scale. The maximum value in each EER column is shaded.

System	Random-target for μ_{AR}	λ_s	λ_c	Development				Evaluation			
				Lazy-EER	Semi-EER	WER	UAR	Lazy-EER	Semi-EER	WER	UAR
Orig.		/	/	7.34	–	1.80	69.08	3.91	–	1.84	71.06
B2		/	/	8.15	4.23	10.44	55.66	5.62	3.72	9.96	53.49
B3		/	/	26.03	16.32	4.31	38.08	19.06	14.24	4.31	35.24
B4		/	/	19.73	15.28	6.16	41.97	14.64	11.13	5.90	42.78
B5		/	/	30.94	20.05	4.91	40.08	25.65	17.05	4.44	38.25
Built-in anon.		0.7	0.7	33.15	24.64	4.18	41.85	28.19	17.78	4.58	40.38
Smooth-21	✓	1.1	2.8	35.09	25.14	4.53	40.38	29.39	21.51	4.30	40.16

Table 3: Track 2 comparison with the official multilingual baselines and Seed-VC built-in anonymization. Privacy is measured by lazy-informed and semi-informed EER (%), linguistic utility by WER (%), and emotion utility by UAR (%). Global denotes the selected configuration of the proposed method.

Lan.	Metric	Development						Evaluation					
		Orig.	BM1	BM2	BM3	Global	Built-in	Orig.	BM1	BM2	BM3	Global	Built-in
fr	Lazy-EER	6.15	8.02	25.43	47.43	21.37	17.70	7.51	9.49	28.48	48.49	25.80	20.29
	Semi-EER	6.15	1.57	15.45	46.11	9.28	8.32	7.51	2.08	15.30	46.25	9.93	9.10
	WER	5.80	12.42	52.17	13.77	21.38	14.93	5.13	9.93	42.24	11.99	18.66	12.97
es	Lazy-EER	8.97	9.31	28.78	48.78	24.71	21.13	4.86	5.76	25.03	48.19	18.89	16.01
	Semi-EER	8.97	0.93	10.74	46.23	8.78	8.87	4.86	1.58	8.74	45.92	7.50	7.29
	WER	3.79	6.15	20.37	10.56	12.27	10.21	3.75	6.21	28.92	10.12	12.22	9.72
en	Lazy-EER	2.24	2.31	9.73	36.68	10.09	8.88	4.75	4.27	11.55	36.87	11.86	10.42
	Semi-EER	2.24	1.19	5.25	22.02	7.02	7.86	4.75	3.40	7.53	20.40	7.88	9.01
	WER	4.57	7.40	14.29	6.71	10.95	11.36	5.46	8.54	15.41	7.73	12.08	11.43
de	Lazy-EER	3.00	5.86	25.25	48.25	20.41	17.92	5.02	8.75	25.72	48.24	22.76	18.90
	Semi-EER	3.00	1.12	10.05	46.54	8.96	8.92	5.02	1.36	10.90	46.05	10.21	10.57
	WER	5.25	7.81	19.28	9.74	15.12	12.21	5.38	9.38	24.40	10.10	18.58	14.21
Avg.	Lazy-EER	5.09	6.38	22.30	45.29	19.15	16.41	5.54	7.07	22.70	45.45	19.83	16.41
	Semi-EER	5.09	1.20	10.37	40.23	8.51	8.50	5.54	2.11	10.62	39.66	8.88	8.99
	WER	4.85	8.45	26.53	10.20	14.93	12.18	4.93	8.52	27.74	9.98	15.38	12.08
emo	UAR	72.09	41.30	27.13	26.24	28.07	27.72	78.90	42.63	29.04	24.84	26.62	27.64

Track 2, we retain the same guidance scales used in Track 1, namely $\lambda_s = 1.2$ and $\lambda_c = 2.9$, and keep all other system settings unchanged. The only change is the random-target sampling strategy: instead of using the fixed 500-utterance pool adopted in Track 1, each random-target utterance is sampled directly from the full LibriTTS train-other-500 dataset. Averaged across languages, Global achieves higher Lazy-EER and Semi-EER than BM1 on both the development and evaluation sets, and higher Lazy-EER and Semi-EER than the built-in system on the development set, showing improved multilingual speaker privacy. However, its privacy performance still falls behind the stronger BM2 and BM3 baselines, especially BM3. In addition, its WER is higher and its UAR is lower than the best-performing baselines, indicating substantial room for improvement in linguistic and emotion preservation. Overall, the proposed configuration achieves a certain multilingual privacy–utility trade-off without additional Track 2 tuning, but further adaptation is required to obtain stronger multilingual performance.

5. Conclusion

We presented a training-free, two-stage voice anonymization method built on Seed-VC v2. The ablation results show that random-target-conditioned AR construction reduces source-

speaker dependence in μ_{AR} , while source-conditioning-aware projected CFG further suppresses source-aligned information carried by the complete token guidance during speech generation. The selected system improves privacy while maintaining competitive linguistic and emotion utility on Track 1. On Track 2, without additional tuning, it achieves a moderate multilingual privacy–utility trade-off: its privacy performance lies between BM1 and BM2, while its WER is better than BM2 but worse than BM1 and BM3. A clear gap therefore remains to the strongest BM3 baseline.

Future work will further adapt the condition-construction and projection-guided generation strategies to multilingual speech, with the goal of achieving a stronger privacy–utility trade-off for multilingual anonymization.

References

- [1] T. Bäckström, “Privacy in speech technology,” *Proceedings of the IEEE*, 2025.
- [2] A. S. Shamsabadi, B. M. L. Srivastava, A. Bellet, N. Vauquier, E. Vincent, M. Maouche, M. Tommasi, and N. Papernot, “Differentially private speaker anonymization,” *Proceedings on Privacy Enhancing Technologies*, vol. 2023, no. 1, Jan. 2023. [Online]. Available: <https://hal.inria.fr/hal-03588932>

- [3] N. Tomashenko, B. M. L. Srivastava, X. Wang, E. Vincent, A. Nautsch, J. Yamagishi, N. Evans, J. Patino, J.-F. Bonastre, P.-G. Noé *et al.*, “Introducing the voiceprivacy initiative,” *arXiv preprint arXiv:2005.01387*, 2020.
- [4] N. Tomashenko, X. Wang, E. Vincent, J. Patino, B. M. L. Srivastava, P.-G. Noé, A. Nautsch, N. Evans, J. Yamagishi, B. O’Brien *et al.*, “The voiceprivacy 2020 challenge: Results and findings,” *Computer Speech & Language*, vol. 74, p. 101362, 2022.
- [5] N. Tomashenko, X. Wang, X. Miao, H. Nourtel, P. Champion, M. Todisco, E. Vincent, N. Evans, J. Yamagishi, and J.-F. Bonastre, “The voiceprivacy 2022 challenge evaluation plan,” 2022. [Online]. Available: <https://arxiv.org/abs/2203.12468>
- [6] M. Panariello, N. Tomashenko, X. Wang, X. Miao, P. Champion, H. Nourtel, M. Todisco, N. Evans, E. Vincent, and J. Yamagishi, “The voiceprivacy 2022 challenge: Progress and perspectives in voice anonymisation,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 3477–3491, 2024.
- [7] N. Tomashenko, X. Miao, P. Champion, S. Meyer, X. Wang, E. Vincent, M. Panariello, N. Evans, J. Yamagishi, and M. Todisco, “The voiceprivacy 2024 challenge evaluation plan,” *arXiv preprint arXiv:2404.02677*, 2024.
- [8] N. Tomashenko, X. Miao, P. Champion, S. Meyer, M. Panariello, X. Wang, N. Evans, E. Vincent, J. Yamagishi, and M. Todisco, “The third voiceprivacy challenge: Preserving emotional expressiveness and linguistic content in voice anonymization,” *Computer Speech & Language*, vol. 100, p. 101988, 2026. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0885230826000513>
- [9] X. Miao, N. Tomashenko, R. Arefeen, S. Meyer, M. Panariello, X. Wang, E. Vincent, N. Evans, J. Yamagishi, and M. Todisco, “The voiceprivacy 2026 challenge evaluation plan,” 2026.
- [10] P. Gupta, G. P. Prajapati, S. Singh, M. R. Kamble, and H. A. Patil, “Design of voice privacy system using linear prediction,” in *Proc. APSIPA ASC*. IEEE, 2020, pp. 543–549.
- [11] S. P. Dubagunta, R. Van Son, and M. M. Doss, “Adjustable deterministic pseudonymisation of speech: Idiap-NKI’s submission to VoicePrivacy 2020 challenge,” URL: <https://www.voiceprivacychallenge.org/docs/Idiap-NKI.pdf>, 2020.
- [12] S. E. McAdams, *Spectral fusion, spectral parsing and the formation of auditory images*. Stanford university, 1984.
- [13] L. Tavi, T. Kinnunen, and R. G. Hautamäki, “Improving speaker de-identification with functional data analysis of f0 trajectories,” *Speech Communication*, vol. 140, pp. 1–10, 2022.
- [14] H. L. Xinyuan, Z. Cai, A. Garg, K. Duh, L. P. García-Perera, S. Khudanpur, N. Andrews, and M. Wiesner, “Hltcoe jhu submission to the voice privacy challenge 2024,” in *Proc. 4th Symposium on Security and Privacy in Speech Communication*, 2024, pp. 61–66.
- [15] F. Fang, X. Wang, J. Yamagishi, I. Echizen, M. Todisco, N. Evans, and J.-F. Bonastre, “Speaker anonymization using x-vector and neural waveform models,” *arXiv preprint arXiv:1905.13561*, 2019.
- [16] S. Meyer, F. Lux, J. Koch, P. Denisov, P. Tilli, and N. T. Vu, “Prosody is not identity: A speaker anonymization approach using prosody cloning,” in *Proc. ICASSP*. IEEE, 2023, pp. 1–5.
- [17] P. Champion, “Anonymizing speech: Evaluating and designing speaker anonymization techniques,” *arXiv preprint arXiv:2308.04455*, 2023.
- [18] Z. Cai, H. L. Xinyuan, A. Garg, L. P. García-Perera, K. Duh, S. Khudanpur, N. Andrews, and M. Wiesner, “Privacy versus emotion preservation trade-offs in emotion-preserving speaker anonymization,” in *Proc. SLT*. IEEE, 2024, pp. 409–414.
- [19] M. Panariello, F. Nespoli, M. Todisco, and N. Evans, “Speaker anonymization using neural audio codec language models,” in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 4725–4729.
- [20] F. Huang, K. Zeng, and W. Zhu, “DiffVC+: Improving Diffusion-based Voice Conversion for Speaker Anonymization,” in *Interspeech 2024*, 2024, pp. 4453–4457.
- [21] Y. Xiao, Y. Zhao, Z. Ma, S. Wang, K. Yu, J. Liao, and X. Chen, “Anonymization, not elimination: Utility-preserved speech anonymization,” *IEEE Transactions on Audio, Speech and Language Processing*, 2026.
- [22] I. R. Ulgen, Z. Cai, N. Andrews, P. Koehn, and B. Sisman, “Diffanon: Diffusion-based prosody control for voice anonymization,” *arXiv preprint arXiv:2604.26281*, 2026.
- [23] S. Liu, “Zero-shot voice conversion with diffusion transformers,” *arXiv preprint arXiv:2411.09943*, 2024.
- [24] S.-g. Lee, W. Ping, B. Ginsburg, B. Catanzaro, and S. Yoon, “Bigvgan: A universal neural vocoder with large-scale training,” *arXiv preprint arXiv:2206.04658*, 2022.
- [25] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, “Librispeech: An asr corpus based on public domain audio books,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.
- [26] H. Zen, V. Dang, R. Clark, Y. Zhang, R. J. Weiss, Y. Jia, Z. Chen, and Y. Wu, “LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech,” in *Interspeech 2019*, 2019, pp. 1526–1530.