CoReSep: Input Composition for Unifying Speech Separation and Target Speaker Extraction with a Pretrained Speech Foundation Model

Anonymous authors
Paper under double-blind review

Abstract

Speech separation (SS) and target speaker extraction (TSE) share the same underlying goal of disentangling overlapping speakers, providing a natural basis for unified modeling. However, existing unified systems either extend acoustically driven separation backbones with speaker-selection mechanisms, tying source attribution primarily to acoustic evidence, or adopt single-stream language-model generation that requires repeated inference for multi-speaker SS without explicit cross-stream coordination. To address these limitations, we introduce CoReSep, a unified SS and TSE system built around two key ideas. First, separation is performed by a pretrained speech foundation model (SFM) itself, whose contextual prior provides high-level constraints on source attribution beyond acoustic evidence alone. Second, a novel input composition mechanism enables joint multi-stream modeling by specifying each task entirely through the input: replicating the mixture representation instantiates multiple jointly decoded streams for SS, whereas prefixing an enrollment representation anchors the single output stream to the target speaker for TSE. In this way, a standard single-stream SFM serves as the unified separator without architectural modification. A conditional flow-matching module then restores acoustic detail by conditioning on both the separated representation and the mixture Mel spectrogram, after which a vocoder reconstructs the waveform. Experimental results demonstrate that CoReSep outperforms previous task-specific and unified systems in linguistic integrity on both SS and TSE, while maintaining competitive perceptual quality. On Libri2Mix, it achieves differential word error rates (dWERs) of 0.85%/3.65% for SS and 1.22%/3.93% for TSE on the Mix-Clean/Mix-Both subsets, respectively. It also maintains its advantage over a strong discriminative SS baseline on three- and four-speaker mixtures and outperforms it on all metrics on EchoSet, whose reverberation and random overlap ratios are closer to real-world conditions. Audio samples are available online.[1]

1 Introduction

Humans effortlessly attend to a single voice in a crowded room, a phenomenon known as the cocktail party effect (Bronkhorst, 2015). Machines approach this through two lines of work: speech separation (SS),[2] which recovers all constituent sources without prior knowledge of their identities (Luo & Mesgarani, 2019; Subakan et al., 2021), and target speaker extraction (TSE), which recovers a designated speaker given auxiliary cues such as an enrollment utterance (Delcroix et al., 2018; Ge et al., 2020). Although both tasks address the same problem of disentangling speakers from overlapping speech, they have evolved under different conditioning assumptions, resulting in distinct formulations and architectures. This division is limiting in practice, since in real-world applications the number of active speakers is often unknown and a target-speaker enrollment may or may not be available, which motivates a single model that accommodates both tasks.

[1] https://anonymous.4open.science/w/CoReSep-Demo-4E0B/

[2] In this work, we focus on monaural blind speech separation.

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.

Existing unified approaches largely build on conventional separation architectures with speaker-selection mechanisms that accommodate task-dependent output structures: a single stream conditioned on a target cue for TSE, and multiple streams for SS, either a fixed number (Ochiai et al., 2019) or a variable number (Saijo et al., 2023; Chetupalli & Habets, 2024; Wang et al., 2025a) estimated by an encoder-decoder-based attractor (EDA) module (Horiguchi et al., 2022). However, these frameworks operate directly on waveforms or time-frequency features and are typically optimized with signal-level objectives such as the scale-invariant signal-to-distortion ratio (SI-SDR) (Roux et al., 2019). These models therefore tend to attribute sources primarily based on acoustic evidence, leaving them susceptible to *source attribution errors*, including interference leakage, target omission, and within-utterance speaker swaps. Such failures have also been reported for task-specific SS and TSE systems built on the same acoustically driven paradigm (Liu et al., 2023; Zhao et al., 2022; Han & Mesgarani, 2023), particularly when competing speakers have similar vocal characteristics, the target cue is unreliable, or the mixture is heavily corrupted by noise.

To mitigate these errors, recent studies have begun to incorporate phonetic or semantic information into SS and TSE, typically through representations from self-supervised speech foundation models (SFMs) such as HuBERT (Hsu et al., 2021) and WavLM (Chen et al., 2022a). Since speech content evolves coherently over time, such information can in principle constrain source attribution where acoustic evidence alone is ambiguous. Some approaches use continuous SFM representations as inputs or conditioning signals (Peng et al., 2024; Asai et al., 2026). Others adopt language models (LMs) over discrete speech tokens, such as SFM-derived tokens and neural codec tokens, and have shown promise for both TSE (Tang et al., 2024; Li et al., 2026a; Wang et al., 2025b) and SS (Li et al., 2026b). The flexible sequence-modeling capability of LMs further enables general-purpose systems such as LLaSE-G1 (Kang et al., 2025) and UniSE (Yan et al., 2026), which support TSE, SS, and other speech front-end tasks within a single model.

However, two limitations remain. First, these LM-based unified systems realize SS through recursive inference, extracting one speaker at a time conditioned on previous outputs. This is susceptible to error propagation across passes and prevents joint modeling of multiple output streams; moreover, these systems have so far been evaluated only on two-speaker mixtures. The capacity of sequence models to decode multiple coordinated streams in a single pass thus remains largely unexplored for unified SS and TSE. Second, SFMs typically serve as feature or token extractors for a downstream separator, leaving their pretrained contextual prior largely outside the separation process. PASE (Rong et al., 2026b) refers to this prior in WavLM as the *phonological prior*, as it arises primarily from contextual modeling of phonetic patterns, and shows it to be an effective high-level constraint for preserving linguistic content in speech enhancement. Extending this prior to SS and TSE, however, is not straightforward. A standard SFM is designed for single-stream sequence modeling, with no explicit mechanism for incorporating target-speaker enrollment in TSE or jointly producing multiple streams in SS. Consequently, adapting the SFM itself would typically require task-specific architectural modifications, limiting direct use of its pretrained prior.

To address these limitations, we introduce **CoReSep**, a unified system for SS and TSE built around a novel *input composition* mechanism to adapt a pretrained SFM for adaptive-stream modeling. Without introducing task-specific modules, CoReSep configures the two tasks entirely through the input: replicating the mixture representation instantiates multiple jointly decoded streams in a single pass for SS, whereas prefixing an enrollment representation anchors the single output stream to the target speaker for TSE. For SS, surplus streams are trained to predict silence, removing the need for an external speaker-counting module. This formulation enables the pretrained SFM itself to serve as the separator, allowing its phonological prior to provide high-level constraints on source attribution beyond acoustic evidence alone. Since the separated SFM representations primarily capture phonetic information (Pasad et al., 2023; Rong et al., 2026b), a conditional flow-matching (CFM) module recovers the fine-grained acoustic details needed for high-fidelity reconstruction. Conditioned on both the separated representation and the mixture Mel spectrogram, the CFM generates the Mel spectrogram of the corresponding source, which a pretrained Mel vocoder then converts into a waveform. Our main contributions are as follows:

- **Input-Composition-Based Unified Modeling**: We introduce a general input-composition mechanism that enables encoder-only sequence models to unify SS and TSE without architectural modification, while supporting adaptive-stream outputs and joint multi-stream modeling.

- **SFM-Based Separation with Phonological Prior**: We leverage a pretrained SFM directly as the separator, bringing its phonological prior into the separation process as a high-level constraint on source attribution, rather than treating the SFM merely as a feature or token extractor.
- **Strong Empirical Performance**: CoReSep achieves the best linguistic integrity among task-specific and unified systems on both tasks, with dWERs of 0.85%/3.65% (SS) and 1.22%/3.93% (TSE) on the Libri2Mix Mix-Clean/Mix-Both subsets. On three- and four-speaker mixtures, a single CoReSep model outperforms TF-GridNet trained for each speaker count in linguistic integrity, and on the more realistic EchoSet, it outperforms TF-GridNet on all SS metrics.

2 Related Work

2.1 Task-Specific Speech Separation and Target Speaker Extraction

Blind SS typically pairs a fixed number of output heads with permutation-invariant training (PIT) (Yu et al., 2017; Kolbæk et al., 2017), with progress driven largely by increasingly capable time-domain and time-frequency backbones (Luo & Mesgarani, 2019; Subakan et al., 2021; Wang et al., 2023; Zhao et al., 2024). Unknown speaker counts are handled by recursive extraction (Takahashi et al., 2019), attractor-based decoding (Lee et al., 2024), or fixed-size outputs trained to emit silence for absent sources (Wisdom et al., 2021). TSE methods differ mainly in how the enrollment cue is injected: speaker-embedding-based methods fuse a fixed-length embedding from a jointly trained or pretrained speaker encoder into a separation backbone (Ge et al., 2020; Hao et al., 2024), whereas speaker-embedding-free methods retain the enrollment as a sequence, attending to it via cross-attention (Zeng & Li, 2025; Wang et al., 2026) or prepending it to the mixture so that conditioning emerges within a single network (Navon et al., 2025; Shimizu et al., 2025).

Across both tasks, most existing systems remain acoustically driven, separating sources in waveform, spectrogram, or acoustic-latent spaces; SFM-based approaches introduce continuous features or discrete tokens to provide higher-level information beyond local acoustics (Tsai et al., 2022; Peng et al., 2024; Asai et al., 2026; Erdogan et al., 2023; Li et al., 2026b; Tang et al., 2024; Li et al., 2026a; Wang et al., 2025b). In these approaches, the SFM serves mainly as a representation extractor, while a downstream model is responsible for separating the mixture in the SFM representation space. However, such mixture representations may deviate from the desired phonetic properties that the downstream model seeks to exploit and may not be easily disentangled into individual sources. PASE (Rong et al., 2026b) instead adapts the SFM itself to directly produce target representations under its phonological prior, outperforming approaches that delegate enhancement to a downstream model. CoReSep extends this idea to SS and TSE through a novel input-composition mechanism that enables both tasks within the same pretrained architecture.

2.2 Unified Speech Separation and Target Speaker Extraction

Conventional unified systems build on a blind separation backbone that outputs a fixed (Ochiai et al., 2019) or estimated (Saijo et al., 2023; Chetupalli & Habets, 2024; Wang et al., 2025a) number of streams for SS, and add a dedicated module that uses the target cue to obtain the target stream for TSE. Wang et al. (2025a) further extend this paradigm to universal sound separation, using text, visual, and sound-tag cues. Like the task-specific systems above, these approaches remain acoustically driven. From the perspective of input formulation, TUSS (Saijo et al., 2025) is structurally closer to CoReSep: it prepends learnable prompts to the encoded mixture and jointly models them with a Transformer. However, its prompts denote source categories rather than speaker identities, and therefore cannot specify an enrolled target speaker for TSE.

LM-based general-purpose systems such as LLaSE-G1 (Kang et al., 2025) and UniSE (Yan et al., 2026) cast both tasks as conditional generation. They naturally support enrollment-conditioned extraction, but perform SS by recursive inference, extracting one speaker per pass conditioned on previously extracted speech, so that the output streams are decoded sequentially rather than jointly. CoReSep departs from both lines: for SS, all streams are decoded jointly in a single forward pass, each from its own copy of the mixture representation, whereas TSE is specified solely by prepending an enrollment representation, without a selection module, category prompt, or task token.

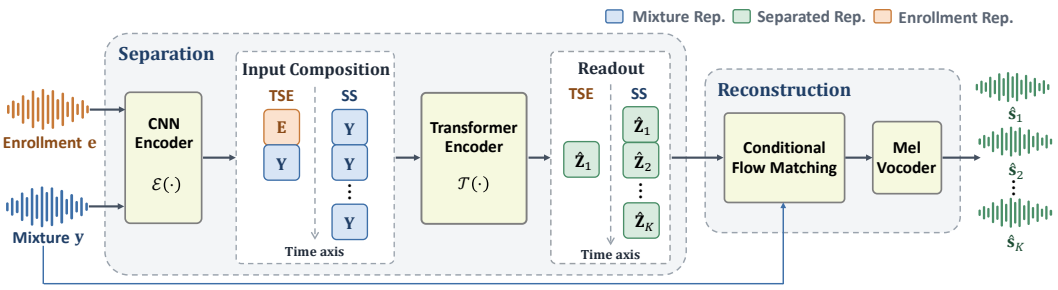

Figure 1: Overall framework of CoReSep. Rep. denotes representation.

3 METHOD

3.1 OVERALL ARCHITECTURE

CoReSep consists of two main components: a *separation* module and a *reconstruction* module, as illustrated in Figure 1. The separation module, termed the separator, is built on the speech foundation model WavLM, which comprises a convolutional neural network (CNN) encoder and a Transformer encoder. The CNN encoder encodes the mixture and, when available, the enrollment utterance into continuous representations, which are then composed into a single input sequence according to the task, either TSE or SS. The Transformer processes this sequence in a single forward pass and performs separation or extraction internally. Its output is subsequently sliced according to the input composition structure to obtain the separated representations. The reconstruction module consists of a CFM module followed by a pretrained Mel vocoder. Conditioned on the separated representations and the mixture Mel spectrogram, the CFM generates the Mel spectrogram of each separated stream, which is then converted into a waveform by the vocoder. Training is performed in two stages: the separator is trained first, after which the CFM module is trained with the separator kept frozen.

3.2 PROBLEM FORMULATION

Let $\mathbf{y} = \sum_{i=1}^{N} \mathbf{s}_i + \mathbf{n} \in \mathbb{R}^L$ denote a mixture of N speech sources and additive noise, and let $\mathbf{e} \in \mathbb{R}^{L_e}$ be an enrollment utterance of a target speaker. We formulate both SS and TSE as estimating a set of output streams $\{\hat{\mathbf{s}}_k\}_{k=1}^{K}$ from $\mathbf{y}$, together with $\mathbf{e}$ in TSE. SS recovers all active sources without prior knowledge of their identities. The model may instantiate $K \geq N$ output streams, where the $K - N$ surplus streams are assigned to silence. TSE recovers only the source corresponding to the speaker specified by $\mathbf{e}$, so $K = 1$ and the single stream is tied to the target speaker.

3.3 INPUT-COMPOSITION-BASED SEPARATION AND EXTRACTION

CNN encoding. To compose the mixture and the enrollment into a single sequence, both must first be expressed in a common representation space. We obtain this space from the CNN encoder $\mathcal{E}(\cdot)$ of the SFM, including its projection to the Transformer dimension D, which encodes both inputs with shared parameters: $\mathbf{Y} = \mathcal{E}(\mathbf{y}) \in \mathbb{R}^{T \times D}$ and $\mathbf{E} = \mathcal{E}(\mathbf{e}) \in \mathbb{R}^{T_e \times D}$. Prior analyses (Chen et al., 2022a; Pasad et al., 2023) have shown that these CNN-encoded representations retain rich acoustic information, including speaker characteristics, so the enrollment can serve as a speaker cue in the same form as the mixture. As a result, $\mathbf{Y}$ and $\mathbf{E}$ can be concatenated directly as input to the Transformer, without a separate speaker encoder or additional representation alignment. The enrollment also remains a sequence of frame-level representations rather than being pooled into a fixed-dimensional speaker embedding, so the Transformer can attend to it frame by frame.

Input composition. A standard pretrained SFM maps one input sequence to one output sequence, whereas SS requires multiple output streams and TSE requires enrollment conditioning. Accommodating these through task-specific output heads or fusion modules would leave multi-stream modeling and speaker conditioning to the added components rather than the SFM itself, potentially limiting the effective use of its pretrained contextual prior. In contrast, our *input composition* specifies the output structure of each task entirely through the input sequence, allowing the SFM itself to pro-

duce the required output streams. It relies only on a generic property of the Transformer encoder: its output is frame-aligned with its input, with each input frame yielding exactly one output frame. Consequently, multiple streams can be packed into a single sequence along the time dimension, where each dedicated segment directly maps to an individual output stream. For SS, we concatenate K copies of $\mathbf{Y}$ to instantiate K output streams; for TSE, we prefix $\mathbf{Y}$ with $\mathbf{E}$ to form a single stream conditioned on the target speaker:

$$\mathbf{X}_{\mathrm{SS}} = [\underbrace{\mathbf{Y}; \ldots; \mathbf{Y}}_{K}] \in \mathbb{R}^{KT \times D}, \qquad \mathbf{X}_{\mathrm{TSE}} = [\,\mathbf{E}; \mathbf{Y}\,] \in \mathbb{R}^{(T_e+T) \times D}, \tag{1}$$

where $[\cdot;\cdot]$ denotes concatenation along the time axis. The two tasks thus differ only in the composed input, and the separator retains exactly the architecture of the pretrained SFM, with no newly initialized component. It can therefore be initialized entirely from the pretrained weights, and a single set of parameters serves both tasks. Moreover, since no parameter is tied to a particular stream, the number of output streams is determined by the input alone.

Transformer modeling and readout. Input composition arranges the streams but does not itself separate them. Separation and extraction are instead performed by the Transformer's contextual modeling over the composed sequence, $\mathbf{H} = \mathcal{T}(\mathbf{X})$, without any explicit separation or speaker-selection head. In TSE, the mixture segment can attend to the preceding enrollment segment, from which it obtains the target-speaker information needed for extraction. In SS, each copy of $\mathbf{Y}$ can attend to all other copies, so each stream is computed from its own copy of the mixture while having access to the other streams. All streams are thus decoded jointly in a single forward pass, rather than being extracted one speaker per pass or read out by multiple heads from a shared encoding of the mixture. Although the copies are identical in content, they occupy distinct positions relative to one another and to the sequence boundaries. Since composition is applied after the CNN encoder, this positional information is introduced only in WavLM's Transformer, through its convolution-based relative position embedding and gated relative position bias, which operate on the composed sequence and thus make the copies distinguishable without external stream identifiers or delimiters. Position alone, however, does not determine which speaker each copy corresponds to; this is learned through the permutation-invariant objective described below. Because the separator is fine-tuned from the pretrained WavLM rather than trained from scratch, it inherits the contextual modeling capacity acquired during pretraining, allowing WavLM's phonological prior to act directly on both tasks and constrain each stream beyond acoustic evidence alone. Finally, since $\mathbf{H}$ is frame-aligned with $\mathbf{X}$, readout reduces to slicing: the separated representation $\hat{\mathbf{Z}}_k \in \mathbb{R}^{T \times D}$ of stream k is taken from the positions occupied by its copy of $\mathbf{Y}$, and the outputs at the enrollment positions are discarded in TSE. Each $\hat{\mathbf{Z}}_k$ is therefore time-aligned with the mixture and can be supervised directly with frame-level targets.

3.4 Stream Assignment and Training Objective

Since the separator outputs representations rather than waveforms, we supervise it in the same representation space, with targets provided by a frozen copy of the pretrained WavLM: for each clean source $\mathbf{s}_i$, the target $\mathbf{Z}_i \in \mathbb{R}^{T \times D}$ is its final-layer representation, matching the layer at which $\hat{\mathbf{Z}}_k$ is produced. In SS, two further issues arise: a mixture may contain fewer active speakers than streams ($N < K$), and the output streams carry no predefined order.

For the former, the target set is padded to K entries with a silence target $\mathbf{Z}_\varnothing$, i.e., $\mathbf{Z}_i = \mathbf{Z}_\varnothing$ for $i = N+1, \ldots, K$, where $\mathbf{Z}_\varnothing$ is obtained by passing an all-zero waveform through the same frozen model. Since $\mathbf{Z}_\varnothing$ lies in the same space as the speech targets, it requires no separate loss formulation. For the latter, we match the K streams to the K targets by utterance-level PIT (Kolbæk et al., 2017), minimizing the mean squared error (MSE) under the best assignment:

$$\mathcal{L} = \min_{\pi} \frac{1}{KTD} \sum_{k=1}^{K} \left\| \hat{\mathbf{Z}}_k - \mathbf{Z}_{\pi(k)} \right\|_F^2, \tag{2}$$

where π ranges over all permutations of $\{1, \ldots, K\}$. In TSE, the single stream is already tied to the enrolled speaker, so the loss reduces to the MSE against that speaker's target without permutation search. Both tasks are therefore trained jointly under a single objective. When the speaker count is unknown, it can be read off from the decoded streams (Appendix A.4).

3.5 RECONSTRUCTION

Although the separator effectively separates the sources in the representation space, these final-layer features predominantly retain phonetic information and lack the low-level acoustic details, such as speaker timbre, needed for faithful waveform reconstruction. The reconstruction module recovers these fine-grained acoustic properties using a CFM module followed by a pretrained Mel vocoder. The CFM module is conditioned on both the separated representation $\hat{\mathbf{Z}}_k$ and the mixture Mel spectrogram, with the former providing the phonetic content of the target stream and the latter supplying the acoustic details needed to preserve speaker characteristics. It predicts the Mel spectrogram of the target stream, which is then converted into a waveform by the vocoder.

In the second training stage, the CFM module is trained on top of the frozen separator with the same objective and masked-infilling scheme as SenSE (Li et al., 2026c) and StuPASE (Rong et al., 2026a), minimizing the MSE between the predicted and target velocities within the masked region. The stream-to-source assignment is determined by applying the PIT matching of Section 3.4 to the frozen separator's outputs and is kept fixed, so no permutation search is required in the CFM loss. Streams assigned to $\mathbf{Z}_\varnothing$ are excluded from CFM training.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Data. Training mixtures are generated on the fly (Zeghidour & Grangier, 2021) using speech from LibriSpeech (Panayotov et al., 2015) (`train-clean-100/360` and `train-other-500`, totaling about 1,000 h). Each interferer is scaled independently to a signal-to-interference ratio (SIR) drawn from $\mathcal{U}[-5, 5]$ dB relative to the target. With a probability of 0.5, WHAM! (Wichern et al., 2019) noise is additionally mixed in at an SNR sampled from $\mathcal{U}[-5, 10]$ dB. Evaluation uses the `mix-clean` and `mix-both` test sets of Libri2Mix and Libri3Mix (Cosentino et al., 2020) under the `min` setting, and a Libri4Mix test set constructed with the official generation scripts extended to four sources. We additionally evaluate on the EchoSet (Xu et al., 2025) test set, whose noise, realistic reverberation, and random overlap ratios make it closer to real-world conditions. All audio is at 16 kHz. Mixing, cropping, and enrollment details are provided in Appendix A.1.

Baselines. We compare CoReSep with three groups of baselines: TSE-specific systems (USEF-TFGridNet (Zeng & Li, 2025), SoloSpeech (Wang et al., 2026), MeanFlow-TSE (Shimizu et al., 2025), LauraTSE (Tang et al., 2025)), SS-specific systems (SepFormer (Subakan et al., 2021), TF-GridNet (Wang et al., 2023), MossFormer2 (Zhao et al., 2024)), and unified systems (UniSE (Yan et al., 2026), LLaSE-G1 (Kang et al., 2025)). USEF-TFGridNet, SepFormer, and TF-GridNet are retrained with our data pipeline; the remaining systems are evaluated with their official checkpoints. Training and inference details are given in Appendix A.2.

Metrics. We evaluate perceptual quality with DNSMOS (Reddy et al., 2021) and UTMOS (Saeki et al., 2022); speaker fidelity with speaker similarity to the clean reference (SpkSim), computed using a WavLM-Large-based ECAPA-TDNN (Chen et al., 2022b); and linguistic integrity with Levenshtein phoneme similarity (LPS) (Pirklbauer et al., 2023), SpeechBERTScore (SBS) (Saeki et al., 2024), and differential word error rate (dWER). We define dWER as the WER between the output transcription and the clean-reference transcription, both obtained using `cohere-transcribe-03-2026` (Mack et al., 2026). Since attribution errors, including interference leakage, target omission, and speaker swapping, can corrupt the output transcription, we use dWER as the primary indicator of attribution errors.

Implementation. The separator is initialized from the official WavLM-Large checkpoint and fully fine-tuned, with each mini-batch assigned to SS or TSE with equal probability. The reconstruction module follows StuPASE (Rong et al., 2026a), with a 16-layer DiT (Peebles & Xie, 2023; Chen et al., 2025) as the CFM backbone and a 12-layer Vocos (Siuzdak, 2024) as the Mel vocoder. All hyperparameters are listed in Appendix A.3, and the computational cost of input composition, which grows with the number of composed streams, is analyzed in Appendix A.5.

Table 1: Comparison with baseline systems on Libri2Mix, Libri3Mix, and Libri4Mix. Best result per column within each mode is **bold**. DNS: DNSMOS; UT: UTMOS.

Method	Mode	Mix-Clean						Mix-Both					
		DNS↑	UT↑	SpkSim↑	LPS↑	SBS↑	dWER↓	DNS↑	UT↑	SpkSim↑	LPS↑	SBS↑	dWER↓
Libri2Mix													
USEF-TFGridNet	TSE	3.74	3.79	0.92	0.95	0.90	3.03	3.61	3.32	0.79	0.88	0.80	6.04
SoloSpeech	TSE	3.84	3.81	0.90	0.91	0.87	5.49	3.72	3.42	0.79	0.85	0.78	11.66
MeanFlow-TSE	TSE	3.69	3.89	**0.94**	0.96	**0.92**	1.78	3.55	3.63	**0.81**	0.90	0.82	5.47
LauraTSE	TSE	3.84	**4.14**	0.82	0.93	0.89	3.37	–	–	–	–	–	–
SepFormer	SS	3.41	3.16	0.88	0.95	0.85	1.67	3.05	2.31	0.68	0.80	0.72	8.13
TF-GridNet	SS	3.68	3.61	**0.93**	**0.97**	0.89	1.10	3.52	2.97	**0.80**	0.88	0.77	4.99
MossFormer2	SS	3.73	3.95	0.91	0.93	0.88	5.58	3.56	3.66	0.76	0.86	0.79	8.94
UniSE	TSE	3.77	**4.14**	0.63	0.82	0.83	16.46	3.76	**4.08**	0.52	0.62	0.72	43.96
	SS	3.77	**4.13**	0.60	0.83	0.83	15.48	3.75	4.03	0.45	0.62	0.71	44.86
LLaSE-G1	TSE	3.51	2.73	0.35	0.65	0.70	27.95	3.43	2.48	0.25	0.45	0.60	58.39
	SS	3.49	2.38	0.28	0.55	0.65	39.79	3.41	2.14	0.19	0.36	0.56	68.61
CoReSep	TSE	**3.89**	3.93	0.86	**0.97**	**0.92**	**1.22**	**3.96**	**4.08**	0.77	**0.94**	**0.88**	**3.93**
	SS	**3.89**	3.94	0.86	**0.97**	**0.92**	**0.85**	**3.96**	**4.08**	0.76	**0.95**	**0.88**	**3.65**
Libri3Mix													
TF-GridNet	SS	3.32	2.68	**0.80**	0.85	0.76	6.55	3.18	2.20	**0.68**	0.69	0.66	17.65
CoReSep	TSE	3.86	3.95	0.75	0.93	0.88	5.50	3.88	4.00	0.67	0.88	0.84	11.83
	SS	**3.86**	**3.95**	0.75	**0.94**	**0.88**	**4.07**	**3.88**	**4.00**	0.65	**0.89**	**0.84**	**10.76**
Libri4Mix													
TF-GridNet	SS	2.99	1.83	**0.65**	0.56	0.60	27.56	2.92	1.62	**0.54**	0.42	0.54	43.19
CoReSep	TSE	3.84	3.90	0.66	0.85	0.83	14.93	3.84	3.90	0.59	0.78	0.79	24.68
	SS	**3.83**	**3.84**	0.62	**0.82**	**0.82**	**18.17**	**3.84**	**3.83**	**0.54**	**0.74**	**0.76**	**29.28**

Table 2: Results on the reverberant EchoSet test set. All systems are evaluated in the SS mode.

Method	DNSMOS↑	UTMOS↑	SpkSim↑	LPS↑	SBS↑	dWER(%)↓
TF-GridNet	3.30	1.81	0.75	0.86	0.73	7.73
CoReSep	**3.76**	**3.90**	**0.81**	**0.93**	**0.89**	**5.74**

4.2 Comparison with Prior Systems

Table 1 summarizes the results on Libri2Mix, Libri3Mix, and Libri4Mix. For Libri2Mix, CoReSep is trained on two-speaker mixtures, whereas for Libri3Mix and Libri4Mix, it is trained on mixtures of two to four speakers. In all cases, the number of speakers is assumed known at inference, so SS instantiates $K = N$ output streams. Models are grouped by supported tasks; we first focus on Libri2Mix to establish an overall comparison with the baselines.

TSE: CoReSep achieves the lowest dWER in both conditions, reducing it by 28%–31% relative to the strongest TSE-specific baseline, MeanFlow-TSE (1.22% vs. 1.78% on Mix-Clean and 3.93% vs. 5.47% on Mix-Both), while matching or surpassing all baselines on LPS and SBS. It also attains the highest DNSMOS in both conditions and remains competitive on UTMOS.

SS: CoReSep likewise achieves the lowest dWER in both conditions, reducing it by 23%–27% relative to the strongest SS baseline, TF-GridNet (0.85% vs. 1.10% and 3.65% vs. 4.99%). This advantage is particularly pronounced under noise, where acoustic baselines suffer steeper drops in linguistic integrity; for instance, the SBS of TF-GridNet drops from 0.89 to 0.77, whereas that of CoReSep drops only from 0.92 to 0.88.

In both tasks, CoReSep trails the strongest acoustic baselines in SpkSim (e.g., 0.86 vs. 0.94 for MeanFlow-TSE and 0.93 for TF-GridNet on Mix-Clean). As analyzed in Appendix A.6, this gap is mainly attributable to the SpkSim ceiling of the reconstruction back-end.

Unified systems: CoReSep shows an even larger advantage over the general-purpose unified systems UniSE and LLaSE-G1, reducing dWER by more than an order of magnitude. Notably, LLaSE-

Table 3: Ablation on unified modeling and the pretrained SFM prior on Libri2Mix. FS: trained from scratch. All outputs are decoded by the layer-24 WavLM vocoder without the CFM back-end.

Model	Mode	Mix-Clean			Mix-Both		
		LPS	SBS	dWER (%)	LPS	SBS	dWER (%)
Separator-TSEonly	TSE	0.96	0.91	1.36	0.94	0.88	4.06
Separator-SSonly	SS	0.96	0.91	1.45	0.94	0.87	4.79
Separator	TSE	0.96	0.91	1.51	0.94	0.88	3.92
	SS	0.96	0.91	1.29	0.95	0.87	3.58
Separator-FS	TSE	0.92	0.87	5.99	0.83	0.80	16.73
	SS	0.93	0.87	4.93	0.83	0.80	16.23

G1 degrades substantially from TSE to SS, whereas CoReSep maintains comparable perceptual quality and speaker fidelity across the two tasks. The results also highlight that perceptual quality alone does not reliably reflect source attribution: for instance, UniSE achieves the highest UTMOS on Mix-Clean despite substantially higher dWER.

Overall, CoReSep achieves consistently stronger linguistic integrity and source attribution on both TSE and SS while maintaining competitive perceptual quality. Its advantage is particularly evident under noise, suggesting greater robustness to degraded acoustic evidence. Moreover, as a single model surpasses task-specific systems in linguistic integrity on both tasks, unified modeling does not come at the expense of task-specific effectiveness.

We next evaluate scalability to larger speaker counts on Libri3Mix and Libri4Mix. Since most baselines lack suitable checkpoints, we retrain TF-GridNet separately for 3 and 4 speakers, whereas a single CoReSep model handles both. While all models degrade with more speakers, CoReSep consistently outperforms TF-GridNet in perceptual quality and linguistic integrity, widening its advantage on LPS, SBS, and UTMOS.

Finally, we evaluate SS on EchoSet (which lacks enrollment), with both CoReSep and TF-GridNet retrained on reverberant mixtures (Appendix A.1). As shown in Table 2, CoReSep outperforms TF-GridNet across all metrics, lowering dWER from 7.73% to 5.74% (26% relative) and improving LPS (0.93 vs. 0.86) and SBS (0.89 vs. 0.73), while UTMOS more than doubles (3.90 vs. 1.81). Notably, CoReSep also achieves higher SpkSim (0.81 vs. 0.75), reversing TF-GridNet's advantage on LibriMix. This confirms that CoReSep's advantages carry over to realistic acoustic conditions.

4.3 Ablation Study

We first examine whether unified modeling compromises the performance of either SS or TSE, and then validate two key design choices: (i) the pretrained SFM prior and (ii) input composition. To isolate the separator itself, all ablations exclude the CFM back-end, and metrics are computed by decoding the separator outputs with a dedicated vocoder trained on layer-24 WavLM representations.

Unified modeling. Table 3 compares the unified Separator with task-specific variants trained only on TSE or SS. Across tasks and conditions, the unified model achieves comparable LPS and SBS while obtaining lower dWER in three of the four settings, demonstrating that joint modeling does not compromise task-specific performance.

Pretrained SFM prior. We compare two variants with the same WavLM-Large architecture: the default Separator, initialized from the pretrained checkpoint to inherit its phonological prior, and Separator-FS, trained from scratch with random initialization. Apart from the learning rate and training length, which are adjusted for training from scratch (Appendix A.3), all other settings are identical. As shown in Table 3, removing the pretrained initialization increases dWER roughly fourfold in both conditions, and the degradation is more severe under noise, where dWER exceeds 16% and LPS drops from 0.94–0.95 to 0.83. This highlights the importance of the phonological prior, particularly under challenging acoustic conditions.

Table 4: Ablation on input composition: dWER (%) on the Mix-Both subsets of Libri2Mix, Libri3Mix, and Libri4Mix. Both composition variants are trained on two- to four-speaker mixtures; the SUPERB-style baseline is trained separately for each speaker count. Decoding follows Table 3.

Architecture	Multi-Stream Implementation	Libri2Mix		Libri3Mix		Libri4Mix	
		TSE	SS	TSE	SS	TSE	SS
WavLM	Input Composition	**5.55**	**3.56**	**11.63**	**10.52**	**23.53**	**28.27**
	Zero-Padded Composition	5.68	6.21	16.49	18.93	35.93	42.95
WavLM+Heads	SUPERB-style	–	5.87	–	21.47	–	39.81

PASE (Rong et al., 2026b) interprets this prior primarily as a high-level constraint on linguistic content, where intact context guides the recovery of corrupted speech. For SS and TSE, however, source attribution also requires acoustic information to distinguish speakers. These results suggest that the pretrained SFM prior operates hierarchically: lower layers preserve acoustic and speaker-related cues, while higher layers provide increasingly phonetic and contextual constraints. The former can support speaker discrimination and enrollment matching, whereas the latter can help maintain a coherent linguistic trajectory within each output stream. In this sense, the same pretrained prior may play a broader role in TSE and SS, extending beyond purely phonetic constraints by integrating acoustic and contextual information across the SFM hierarchy.

Input composition. We compare input composition with two alternative multi-stream implementations: zero-padded composition, which retains the composed sequence length but replaces the additional mixture copies with zeros, and a SUPERB-style baseline (Tsai et al., 2022), which keeps a single-stream mixture input and reads out the speaker streams through multiple output heads. As its number of heads is fixed, the SUPERB-style baseline is trained separately for each speaker count, whereas both composition variants are trained once on two- to four-speaker mixtures (details in Appendix A.3).

As shown in Table 4, input composition consistently achieves the best SS performance, with its absolute advantage growing on Libri3Mix and Libri4Mix. Zero-padded composition is close on Libri2Mix TSE but already lags on SS, and the gap widens with more speakers. A likely reason is that only the first segment contains the mixture, so later streams must retrieve source information over increasingly long distances. The SUPERB-style baseline also degrades markedly with more speakers despite per-count training, possibly because the shared encoder produces a mixture representation that is not fully disentangled, leaving separation largely to lightweight output heads with limited capacity and no explicit cross-stream interaction. Input composition avoids both issues by giving each stream its own copy of the mixture, allowing all streams to interact within the shared Transformer and leveraging its full modeling capacity for separation. It also consistently outperforms zero-padded composition on TSE. Since the TSE input is identical under both schemes, this gap likely stems from the weaker SS formulation degrading the jointly trained separator.

5 Conclusion

We presented CoReSep, which unifies SS and TSE by using a pretrained SFM itself as the separator, so that its phonological prior constrains source attribution. Input composition specifies each task without architectural modification: replicating the mixture representation enables joint decoding of all SS streams, and prefixing an enrollment representation anchors the TSE output to the target speaker. A CFM-based reconstruction module then restores acoustic detail. CoReSep achieves the lowest dWER among all compared systems in both tasks on LibriMix and outperforms TF-GridNet on all SS metrics on the noisy, reverberant EchoSet. Ablations confirm that unified training compromises neither task, that pretrained initialization substantially reduces source attribution errors, especially under noise, and that input composition outperforms alternative multi-stream implementations, with larger absolute gains for more speakers. Remaining limitations are that errors and computational cost grow with the number of speakers, and that speaker similarity on LibriMix, limited mainly by the reconstruction module, trails the strongest acoustically driven systems.

AI USE STATEMENT

We did not use generative AI tools for any of the tasks with required disclosure. AI tools were used only sparingly for minor language refinement and editorial improvements, and all such edits were reviewed by the authors. We take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

All datasets and the pretrained SFM model (WavLM) used in this work are publicly available. Detailed experimental settings, including hyperparameters and implementation details, are provided in Section 4.1 and Appendix.

REFERENCES

Kohei Asai, Wataru Nakata, Yuki Saito, and Hiroshi Saruwatari. Geneses: Unified generative speech enhancement and separation. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 22207–22211, 2026. doi: 10.1109/ICASSP55912.2026.11464549.

Adelbert W. Bronkhorst. The cocktail-party problem revisited: early processing and selection of multi-talker speech. *Attention, Perception, & Psychophysics*, 77(5):1465–1487, April 2015. ISSN 1943-393X. doi: 10.3758/s13414-015-0882-9. URL http://dx.doi.org/10.3758/s13414-015-0882-9.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518, 2022a.

Yushen Chen, Zhikang Niu, Ziyang Ma, Keqi Deng, Chunhui Wang, Jian Zhao, Kai Yu, and Xie Chen. F5-TTS: A fairytaler that fakes fluent and faithful speech with flow matching. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 6255–6271, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.313. URL https://aclanthology.org/2025.acl-long.313/.

Zhengyang Chen, Sanyuan Chen, Yu Wu, Yao Qian, Chengyi Wang, Shujie Liu, Yanmin Qian, and Michael Zeng. Large-scale self-supervised speech representation learning for automatic speaker verification. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6147–6151, 2022b. doi: 10.1109/ICASSP43922.2022.9747814.

Srikanth Raj Chetupalli and Emanuël A. P. Habets. A unified approach to speaker separation and target speaker extraction using encoder-decoder based attractors. In *2024 18th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 190–194, 2024. doi: 10.1109/IWAENC61483.2024.10694625.

Joris Cosentino, Manuel Pariente, Samuele Cornell, Antoine Deleforge, and Emmanuel Vincent. Librimix: An open-source dataset for generalizable speech separation, 2020. URL https://arxiv.org/abs/2005.11262.

Marc Delcroix, Katerina Zmolikova, Keisuke Kinoshita, Atsunori Ogawa, and Tomohiro Nakatani. Single channel target speaker extraction and recognition with speaker beam. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5554–5558, 2018. doi: 10.1109/ICASSP.2018.8462661.

Hakan Erdogan, Scott Wisdom, Xuankai Chang, Zalán Borsos, Marco Tagliasacchi, Neil Zeghidour, and John R. Hershey. TokenSplit: Using Discrete Speech Representations for Direct, Refined, and Transcript-Conditioned Speech Separation and Recognition. In *Interspeech 2023*, pp. 3462–3466, 2023. doi: 10.21437/Interspeech.2023-2069.

Meng Ge, Chenglin Xu, Longbiao Wang, Eng Siong Chng, Jianwu Dang, and Haizhou Li. SpEx+: A Complete Time Domain Speaker Extraction Network. In *Interspeech 2020*, pp. 1406–1410, 2020. doi: 10.21437/Interspeech.2020-1397.

Cong Han and Nima Mesgarani. Online binaural speech separation of moving speakers with a wavesplit network. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10095695.

Fengyuan Hao, Xiaodong Li, and Chengshi Zheng. X-tf-gridnet: A time–frequency domain target speaker extraction network with adaptive speaker embedding fusion. *Information Fusion*, 112:102550, 2024. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2024.102550. URL https://www.sciencedirect.com/science/article/pii/S1566253524003282.

Shota Horiguchi, Yusuke Fujita, Shinji Watanabe, Yawen Xue, and Paola García. Encoder-decoder based attractors for end-to-end neural diarization. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:1493–1507, 2022. doi: 10.1109/TASLP.2022.3162080.

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.

Boyi Kang, Xinfa Zhu, Zihan Zhang, Zhen Ye, Mingshuai Liu, Ziqian Wang, Yike Zhu, Guobin Ma, Jun Chen, Longshuai Xiao, Chao Weng, Wei Xue, and Lei Xie. Llase-g1: Incentivizing generalization capability for llama-based speech enhancement, 2025. URL https://arxiv.org/abs/2503.00493.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, Michael L. Seltzer, and Sanjeev Khudanpur. A study on data augmentation of reverberant speech for robust speech recognition. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5220–5224, 2017. doi: 10.1109/ICASSP.2017.7953152.

Morten Kolbæk, Dong Yu, Zheng-Hua Tan, and Jesper Jensen. Multitalker speech separation with utterance-level permutation invariant training of deep recurrent neural networks. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(10):1901–1913, 2017. doi: 10.1109/TASLP.2017.2726762.

Younglo Lee, Shukjae Choi, Byeong-Yeol Kim, Zhong-Qiu Wang, and Shinji Watanabe. Boosting unknown-number speaker separation with transformer decoder-based attractor. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 446–450, 2024. doi: 10.1109/ICASSP48485.2024.10446032.

Haoyang Li, Xuyi Zhuang, Azmat Adnan, Ye Ni, Wei Rao, Shreyas Gopal, Eng Siong Chng, Boon Siew Han, and Yuanjin Zheng. Gentse: Enhancing target speaker extraction via a coarse-to-fine generative language model, 2026a. URL https://arxiv.org/abs/2512.20978.

Tianhua Li, Chenda Li, Wei Wang, Xin Zhou, Xihui Chen, Jianqing Gao, and Yanmin Qian. Slm-ss: Speech language model for generative speech separation. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 15432–15436, 2026b. doi: 10.1109/ICASSP55912.2026.11463173.

Xingchen Li, Hanke Xie, Ziqian Wang, Zihan Zhang, Longshuai Xiao, Shuai Wang, and Lei Xie. Sense: Semantic-aware high-fidelity universal speech enhancement, 2026c. URL https://arxiv.org/abs/2509.24708.

Kai Liu, Ziqing Du, Xucheng Wan, and Huan Zhou. X-sepformer: End-to-end speaker extraction network with explicit optimization on speaker confusion. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023. doi: 10.1109/ICASSP49357.2023.10095609.

Yi Luo and Nima Mesgarani. Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 27(8):1256–1266, 2019. doi: 10.1109/TASLP.2019.2915167.

Julian Mack, Ekagra Ranjan, Walter Beller-Morales, Bharat Venkitesh, and Pierre Richemond. cohere-transcribe-03-2026 (revision d96e814), 2026. URL https://huggingface.co/CohereLabs/cohere-transcribe-03-2026.

Aviv Navon, Aviv Shamsian, Yael Segal-Feldman, Neta Glazer, Gil Hetz, and Joseph Keshet. FlowTSE: Target Speaker Extraction with Flow Matching. In *Interspeech 2025*, pp. 2965–2969, 2025. doi: 10.21437/Interspeech.2025-970.

Tsubasa Ochiai, Marc Delcroix, Keisuke Kinoshita, Atsunori Ogawa, and Tomohiro Nakatani. A unified framework for neural speech separation and extraction. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6975–6979, 2019. doi: 10.1109/ICASSP.2019.8683448.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015. doi: 10.1109/ICASSP.2015.7178964.

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

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182, 2023. doi: 10.1109/ICCV51070.2023.00387.

Junyi Peng, Marc Delcroix, Tsubasa Ochiai, Oldřich Plchot, Shoko Araki, and Jan Černocký. Target speech extraction with pre-trained self-supervised learning models. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 10421–10425, 2024. doi: 10.1109/ICASSP48485.2024.10448315.

Jan Pirklbauer, Marvin Sach, Kristoff Fluyt, Wouter Tirry, Wafaa Wardah, Sebastian Moeller, and Tim Fingscheidt. Evaluation metrics for generative speech enhancement methods: Issues and perspectives. In *Speech Communication; 15th ITG Conference*, pp. 265–269, 2023. doi: 10.30420/456164052.

Chandan K A Reddy, Vishak Gopal, and Ross Cutler. Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6493–6497, 2021. doi: 10.1109/ICASSP39728.2021.9414878.

Xiaobin Rong, Jun Gao, Zheng Wang, Mansur Yesilbursa, Kamil Wojcicki, and Jing Lu. Stupase: Towards low-hallucination studio-quality generative speech enhancement, 2026a. URL https://arxiv.org/abs/2603.09234.

Xiaobin Rong, Qinwen Hu, Mansur Yesilbursa, Kamil Wojcicki, and Jing Lu. PASE: Leveraging the Phonological Prior of WavLM for Low-Hallucination Generative Speech Enhancement. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(39):32826–32834, Mar. 2026b. doi: 10.1609/aaai.v40i39.40562.

Jonathan Le Roux, Scott Wisdom, Hakan Erdogan, and John R. Hershey. Sdr – half-baked or well done? In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 626–630, 2019. doi: 10.1109/ICASSP.2019.8683855.

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

Takaaki Saeki, Soumi Maiti, Shinnosuke Takamichi, Shinji Watanabe, and Hiroshi Saruwatari. SpeechBERTScore: Reference-Aware Automatic Evaluation of Speech Generation Leveraging NLP Evaluation Metrics. In *Interspeech 2024*, pp. 4943–4947, 2024. doi: 10.21437/Interspeech.2024-1508.

Kohei Saijo, Wangyou Zhang, Zhong-Qiu Wang, Shinji Watanabe, Tetsunori Kobayashi, and Tetsuji Ogawa. A single speech enhancement model unifying dereverberation, denoising, speaker counting, separation, and extraction. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–6, 2023. doi: 10.1109/ASRU57964.2023.10389724.

Kohei Saijo, Janek Ebbers, François G. Germain, Gordon Wichern, and Jonathan Le Roux. Task-aware unified source separation. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10887819.

Riki Shimizu, Xilin Jiang, and Nima Mesgarani. Meanflow-tse: One-step generative target speaker extraction with mean flow, 2025. URL https://arxiv.org/abs/2512.18572.

Hubert Siuzdak. Vocos: Closing the gap between time-domain and fourier-based neural vocoders for high-quality audio synthesis, 2024. URL https://arxiv.org/abs/2306.00814.

Cem Subakan, Mirco Ravanelli, Samuele Cornell, Mirko Bronzi, and Jianyuan Zhong. Attention is all you need in speech separation. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 21–25, 2021. doi: 10.1109/ICASSP39728.2021.9413901.

Naoya Takahashi, Sudarsanam Parthasaarathy, Nabarun Goswami, and Yuki Mitsufuji. Recursive Speech Separation for Unknown Number of Speakers. In *Interspeech 2019*, pp. 1348–1352, 2019. doi: 10.21437/Interspeech.2019-1550.

Beilong Tang, Bang Zeng, and Ming Li. Tselm: Target speaker extraction using discrete tokens and language models, 2024. URL https://arxiv.org/abs/2409.07841.

Beilong Tang, Bang Zeng, and Ming Li. Lauratse: Target speaker extraction using auto-regressive decoder-only language models. In *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–8, 2025. doi: 10.1109/ASRU65441.2025.11434620.

Hsiang-Sheng Tsai, Heng-Jui Chang, Wen-Chin Huang, Zili Huang, Kushal Lakhotia, Shu-wen Yang, Shuyan Dong, Andy Liu, Cheng-I Lai, Jiatong Shi, Xuankai Chang, Phil Hall, Hsuan-Jui Chen, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung-yi Lee. SUPERB-SG: Enhanced speech processing universal PERformance benchmark for semantic and generative capabilities. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8479–8492, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.acl-long.580. URL https://aclanthology.org/2022.acl-long.580/.

Helin Wang, Jiarui Hai, Dongchao Yang, Chen Chen, Kai Li, Junyi Peng, Thomas Thebaud, Laureano Moro-Velázquez, Jesús Villalba, and Najim Dehak. Solospeech: Enhancing intelligibility and quality in target speech extraction through a cascaded generative pipeline. *IEEE Transactions on Audio, Speech and Language Processing*, 34:2408–2420, 2026. doi: 10.1109/TASLPRO.2026.3682073.

Hongyu Wang, Chenda Li, Xin Zhou, Shuai Wang, and Yanmin Qian. Use: A unified model for universal sound separation and extraction, 2025a. URL https://arxiv.org/abs/2512.21215.

Yuancheng Wang, Jiachen Zheng, Junan Zhang, Xueyao Zhang, Huan Liao, and Zhizheng Wu. Metis: A foundation speech generation model with masked generative pre-training. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 131615–131645. Curran Associates, Inc., 2025b. doi: 10.52202/085713-4381. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/be8987a75afeeadfa65007cf7ee25de0-Paper-Conference.pdf.

Zhong-Qiu Wang, Samuele Cornell, Shukjae Choi, Younglo Lee, Byeong-Yeol Kim, and Shinji Watanabe. Tf-gridnet: Integrating full- and sub-band modeling for speech separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 31:3221–3236, 2023. doi: 10.1109/TASLP.2023.3304482.

Gordon Wichern, Joe Antognini, Michael Flynn, Licheng Richard Zhu, Emmett McQuinn, Dwight Crow, Ethan Manilow, and Jonathan Le Roux. WHAM!: Extending Speech Separation to Noisy Environments. In *Interspeech 2019*, pp. 1368–1372, 2019. doi: 10.21437/Interspeech.2019-2821.

Scott Wisdom, Hakan Erdogan, Daniel P. W. Ellis, Romain Serizel, Nicolas Turpault, Eduardo Fonseca, Justin Salamon, Prem Seetharaman, and John R. Hershey. What's all the fuss about free universal sound separation data? In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 186–190, 2021. doi: 10.1109/ICASSP39728.2021.9414774.

Mohan Xu, Kai Li, Guo Chen, and Xiaolin Hu. Tiger: Time-frequency interleaved gain extraction and reconstruction for efficient speech separation. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 70205–70222, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/af790b7ae573771689438bbcfc5933fe-Paper-Conference.pdf.

Haoyin Yan, Chengwei Liu, Shaofei Xue, Xiaotao Liang, Yinghao Liu, Yuxiang Kong, and Zheng Xue. Unise: A unified framework for decoder-only autoregressive lm-based speech enhancement, 2026. URL https://arxiv.org/abs/2510.20441.

Dong Yu, Morten Kolbæk, Zheng-Hua Tan, and Jesper Jensen. Permutation invariant training of deep models for speaker-independent multi-talker speech separation. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 241–245, 2017. doi: 10.1109/ICASSP.2017.7952154.

Neil Zeghidour and David Grangier. Wavesplit: End-to-end speech separation by speaker clustering. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2840–2849, 2021. doi: 10.1109/TASLP.2021.3099291.

Bang Zeng and Ming Li. Usef-tse: Universal speaker embedding free target speaker extraction. *IEEE Transactions on Audio, Speech and Language Processing*, 33:2110–2124, 2025. doi: 10.1109/TASLPRO.2025.3572756.

Shengkui Zhao, Yukun Ma, Chongjia Ni, Chong Zhang, Hao Wang, Trung Hieu Nguyen, Kun Zhou, Jia Qi Yip, Dianwen Ng, and Bin Ma. Mossformer2: Combining transformer and rnn-free recurrent network for enhanced time-domain monaural speech separation. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 10356–10360, 2024. doi: 10.1109/ICASSP48485.2024.10445985.

Zifeng Zhao, Dongchao Yang, Rongzhi Gu, Haoran Zhang, and Yuexian Zou. Target Confusion in End-to-end Speaker Extraction: Analysis and Approaches. In *Interspeech 2022*, pp. 5333–5337, 2022. doi: 10.21437/Interspeech.2022-176.

A Appendix

A.1 Data Pipeline

Training mixtures. All training mixtures are synthesized on the fly with dynamic mixing (Zeghidour & Grangier, 2021), using speech from the `train-clean-100`, `train-clean-360`, and `train-other-500` subsets of LibriSpeech (Panayotov et al., 2015) (about 1,000 h in total) and noise from WHAM! (Wichern et al., 2019). For each mixture, one source is designated as the target. It is the speaker to be extracted in TSE, and in both tasks the interferer levels are set relative to it. Its enrollment is a different utterance of the same speaker. The remaining sources are drawn from distinct other speakers, so no two sources in a mixture share a speaker. No loudness normalization is applied. The target is kept at its native level, and each interferer is scaled, independently of the others, to a signal-to-interference ratio (SIR) drawn from $\mathcal{U}[-5, 5]$ dB relative to the target. With more than two speakers, the SIR is thus defined pairwise against the target rather than against the sum of the other interferers. With probability 0.5, a noise clip is then added at a signal-to-noise ratio (SNR) drawn from $\mathcal{U}[-5, 10]$ dB relative to the speech mixture. The noise is excluded from every reference, so all output streams are trained to suppress it. Finally, the mixture and all sources are rescaled by a common factor such that the largest absolute sample among them equals a peak drawn from $\mathcal{U}[0.5, 0.9]$. Since the factor is shared, all SIR and SNR relations are preserved, and each scaled source remains exactly its component in the mixture, serving directly as the clean reference. Since the peak is taken jointly, no signal clips. The enrollment is rescaled independently to a peak drawn from the same range, so its level carries no information about the level of the target in the mixture.

Reverberant training mixtures. The TF-GridNet and CoReSep models evaluated on EchoSet (Table 2) are trained separately on reverberant mixtures, using room impulse responses (RIRs) from OpenSLR 26 and OpenSLR 28 (Ko et al., 2017). These mixtures follow the pipeline above with one additional step before level scaling: with probability 0.5, every source in the mixture is convolved with a different RIR drawn from the same room, so that the sources share a room but differ in their acoustic paths. SIR scaling, noise addition (again with probability 0.5), and peak rescaling are then applied as described above, with the SIR computed on the reverberant source images. The references are the dry source signals, scaled by the same gains as their reverberant images, so both models are trained to jointly separate and dereverberate; for CoReSep, the teacher representations are likewise computed from the dry sources. When reverberation is applied, the enrollment is convolved with the same RIR as the target.

Segment lengths. The mixture length is drawn from $\mathcal{U}[4, 8]$ s once per batch and shared by all items in the batch. The composed sequences in a batch therefore have equal length and require no batch-level padding or attention masks. Utterances longer than this length are randomly cropped. Shorter ones are zero-padded, with the padding split randomly between the two ends. The enrollment has a fixed length of 3 s, independent of the mixture length, and is cropped or padded in the same way.

Speaker count and stream configuration. CoReSep is trained in two configurations. In the two-speaker configuration, every mixture contains $N = 2$ speakers; in the multi-speaker configuration, N is drawn uniformly from $\{2, 3, 4\}$. The task is chosen after the mixtures are generated: each batch is assigned to SS or TSE with equal probability. Both tasks thus share the same speaker-count distribution, and a TSE mixture contains $N - 1$ interferers. For TSE, a single stream is instantiated ($K = 1$). For SS, the two-speaker configuration uses $K = N = 2$. In the multi-speaker configuration, the speaker count is treated as known with probability 0.75, in which case $K = N$; otherwise it is treated as unknown, and $K = 4$ regardless of N. When $K > N$, the reference sources of the $K - N$ surplus streams are all-zero waveforms, whose representations under the frozen WavLM give the silence target $\mathbf{Z}_\varnothing$ (Section 3.4). Unless stated otherwise, results on Libri2Mix use the two-speaker configuration, and results on Libri3Mix and Libri4Mix use the multi-speaker configuration.

Evaluation sets. Evaluation uses the `mix-clean` and `mix-both` test sets of Libri2Mix and Libri3Mix (Cosentino et al., 2020) under the `min` setting. As no official four-speaker version exists, we construct a Libri4Mix test set by extending the official LibriMix generation scripts to four

sources. It keeps the same source selection from `test-clean`, the same loudness normalization and noise addition, and the same `min`-mode truncation to the shortest source. For all three speaker counts, the `mix-clean` and `mix-both` test sets each contain 3,000 mixtures.

Train–test separation. No LibriMix data is used for training. All training mixtures are synthesized by the pipeline above from the LibriSpeech training subsets, whereas all evaluation sets are built from `test-clean`. Since LibriSpeech training and test speakers are disjoint, no evaluation speaker is seen during training. Likewise, training and evaluation draw WHAM! noise from its disjoint official training and test splits, respectively.

A.2 Baselines and Inference Protocols

Retrained systems. Three baselines are retrained from scratch on the mixtures described in Appendix A.1, so that their comparison with CoReSep is not confounded by differences in training data: USEF-TFGridNet (Zeng & Li, 2025) for TSE, and SepFormer (Subakan et al., 2021) and TF-GridNet (Wang et al., 2023) for SS. Apart from the training data, each system follows its official implementation and configuration without modification, using the ESPnet implementation for TF-GridNet, and all are trained to convergence. TF-GridNet produces a fixed number of outputs, so it is trained separately for two, three, and four speakers, and the speaker-matched model is used on each test set. SepFormer and USEF-TFGridNet are trained in the two-speaker configuration only; both are therefore evaluated only on Libri2Mix.

Released checkpoints. All other systems are evaluated with their officially released checkpoints: SoloSpeech[3], MeanFlow-TSE[4], LauraTSE[5], MossFormer2[6], UniSE[7], and LLaSE-G1[8]; their training configurations are given in the corresponding publications. Except for MossFormer2, whose training data are not publicly documented, all of them are trained on corpora that include LibriSpeech or LibriMix, so the LibriMix test sets are in-domain for them. Several are trained on substantially more data than the 1,000 h used in this work.

Inference protocols. For TSE, each speaker in a test mixture serves in turn as the target, so a test set of N-speaker mixtures yields $3{,}000N$ extraction trials. For all TSE systems, including CoReSep, the enrollment is a randomly selected utterance of the target speaker other than the one contributing to the mixture, randomly cropped to a fixed length. The selection is governed by a fixed random seed, so all systems receive the same enrollment utterance for a given trial. To avoid a train–test mismatch, the crop length matches the enrollment length each system is trained with, which is 3 s for CoReSep. For SS, the output streams of every system, including CoReSep, are aligned to the references by permutation-invariant matching. The matching criterion is SI-SDR for systems that estimate waveforms directly. For systems whose outputs are resynthesized by a vocoder or codec decoder, where SI-SDR is not meaningful, matching is performed in the space of each system's training targets, e.g., the final-layer WavLM representations for CoReSep. All baselines otherwise use the inference settings of their original publications, including the number of sampling steps and the guidance settings of the generative systems. Two cases require comment. The released LauraTSE model does not accept noisy input, so no `mix-both` results are reported for it. UniSE and LLaSE-G1 recover multiple speakers through cascaded inference passes, each conditioned on the previously extracted speech. We follow their released scripts and, consistent with their original evaluations, report them on Libri2Mix only.

A.3 Implementation Details

All modules are trained on four NVIDIA RTX 5880 Ada GPUs with the same learning-rate schedule: a linear warm-up over the first 10% of training steps, followed by cosine annealing to 1×10^{-6}. Batch sizes below are per GPU.

[3] https://huggingface.co/OpenSound/SoloSpeech-models
[4] https://github.com/rikishimizu/MeanFlow-TSE
[5] https://github.com/Beilong-Tang/lauraTSE_code
[6] https://huggingface.co/alibabasglab/MossFormer2_SS_16K
[7] https://github.com/alibaba/unified-audio/tree/main/QuarkAudio-UniSE
[8] https://github.com/Kevin-naticl/LLaSE-G1

Separator. The convolutional encoder and the Transformer encoder are both initialized from the official WavLM-Large checkpoint, and all parameters are updated during training. The targets are produced by a separate, frozen instance of the same checkpoint and are layer-normalized over the joint time and feature axes of each utterance. The separator outputs and the vocoder inputs are normalized in the same way. The separator is trained for 200k steps with AdamW and a peak learning rate of 5×10^{-5}. The batch size is 8 in the two-speaker configuration and 3 in the multi-speaker configuration, shared by SS and TSE batches.

Conditional flow matching. The CFM backbone is a DiT (Peebles & Xie, 2023) with the modifications of Chen et al. (2025). It has 16 layers, 16 attention heads, a hidden dimension of 1024, and a feed-forward dimension of 2048. The 1024-dimensional separated representation $\hat{\mathbf{Z}}_k$ is linearly projected to 512 dimensions before entering the backbone. Following the masked speech-infilling scheme of Li et al. (2026c) and Rong et al. (2026a), the masking ratio is drawn from $\mathcal{U}[0.7, 1.0]$ for the Mel spectrogram of the target source and from $\mathcal{U}[0.5, 1.0]$ for that of the mixture. The module is trained on the outputs of the frozen separator for 100k steps, with a peak learning rate of 2×10^{-4} and a batch size of 16. At inference, the complete Mel spectrogram is generated with the Euler solver in 8 steps, following Li et al. (2026c), using classifier-free guidance with a strength of 0.5.

Vocoders. Log-Mel spectrograms have 100 bins and are computed with a window and FFT size of 1280 samples and a hop size of 320 samples, i.e., at 50 Hz. The Mel vocoder is a 12-layer Vocos (Siuzdak, 2024), trained on the same data as Rong et al. (2026a), which does not include LibriSpeech, for 200k steps with a peak learning rate of 2×10^{-4} and a batch size of 60 one-second segments. The L24 vocoder, used for representation-level evaluation, is identical except that its input is the layer-24 representation of the frozen WavLM, which has the same 50 Hz frame rate. It is trained on the same data for the same number of steps and never observes the mixture.

Ablation variants. All variants use the same targets and optimization settings as the separator unless stated otherwise.

- **Zero-padded composition** keeps one mixture segment and sets the other $K - 1$ segments of the encoded sequence to zero, i.e., $\mathbf{X}_{\mathrm{SS}} = [\mathbf{Y}; \mathbf{0}; \ldots; \mathbf{0}]$, so sequence length, parameter count, and compute are unchanged. The TSE composition $[\mathbf{E}; \mathbf{Y}]$ is also unchanged, so its TSE results differ from those of the separator only through joint training with the modified SS composition.
- **The SUPERB-style baseline** encodes the mixture once and decodes all streams from the shared encoding with K parallel output heads, each a three-layer LSTM with 256 hidden units followed by a linear projection to the D-dimensional target space. Its encoder is initialized from the same checkpoint and trained jointly with the heads. Besides where the streams are formed, it differs from the separator in being trained on SS alone, with one model per speaker count.
- **Separator-FS** uses the same architecture and input composition as the separator but random initialization instead of the pretrained WavLM weights. To account for training from scratch, its peak learning rate is raised to 5×10^{-4} and training is extended to 300k steps; the targets are unchanged.
- **Separator-TSEonly** and **Separator-SSonly** are each trained on a single task under the same recipe, with the same number of optimization steps and the same mixture distribution as the jointly trained separator.

A.4 Speaker Counting

In the main experiments, SS instantiates $K = N$ streams, so the speaker count must be given at inference. TSE needs no count, since its composition $[\mathbf{E}; \mathbf{Y}]$ is independent of the number of interferers. Here we remove this requirement for SS. We run the separator trained in the multi-speaker configuration (Appendix A.1) with $K = 4$ on every test mixture and estimate N from the decoded streams, relying on the surplus streams being trained towards the silence target $\mathbf{Z}_{\varnothing}$ (Section 3.4). All results are on the `mix-both` sets.

Table 5: SS with unknown speaker count on the `mix-both` test sets (3,000 mixtures each), using the separator trained in the multi-speaker configuration with $K = 4$ and $\tau = 0.1$. Over- and under-estimation are the percentages of mixtures with $\hat{N} > N$ and $\hat{N} < N$; over-estimation is impossible on Libri4Mix. dWER is computed from the separator outputs decoded by the L24 vocoder, with the known-N row using the same model with $K = N$, and is not comparable with Table 1.

	Libri2Mix	Libri3Mix	Libri4Mix
Accuracy (%)	98.40	96.43	99.47
Over-estimation (%)	1.60	3.50	—
Under-estimation (%)	0.00	0.07	0.53
dWER (%), known N ($K = N$)	3.56	10.52	28.27
dWER (%), unknown N ($K = 4$)	4.69	10.65	28.27

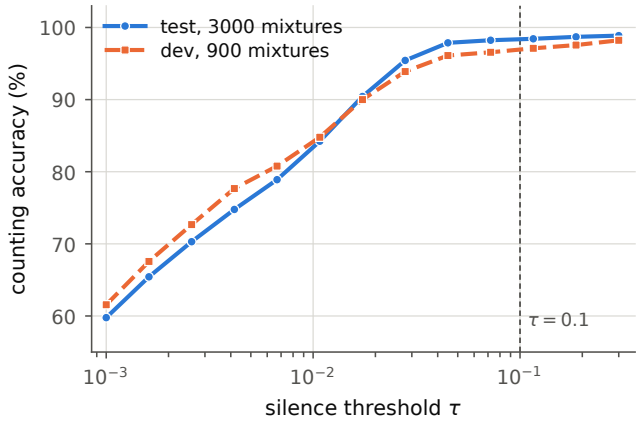

Figure 2: Counting accuracy on the Libri2Mix development and test sets as a function of τ ($K = 4$), evaluated at 13 logarithmically spaced values in $[10^{-3}, 0.3]$ and at the operating point $\tau = 0.1$ (dashed line).

Estimating the speaker count. A stream is declared active if its representation deviates sufficiently from $\mathbf{Z}_\varnothing$:

$$s_j = \frac{1}{TD}\left\|\hat{\mathbf{Z}}_j - \mathbf{Z}_\varnothing\right\|_F^2, \qquad \hat{N} = \left|\{\, j : s_j > \tau \,\}\right|, \tag{3}$$

where $\hat{\mathbf{Z}}_j$ and $\mathbf{Z}_\varnothing$ are layer-normalized over the joint (T, D) axes. Under this normalization, $s_j = 2(1 - \rho_j)$, where ρ_j is the Pearson correlation between the two representations over all TD entries. A surplus stream may carry some residual content and still be treated as silent, so τ need not be tight on that side; it only needs to leave a margin below the statistics of active streams. On the Libri2Mix development set, the smallest statistic of any active stream is 0.34, and we fix $\tau = 0.1$, i.e., a stream is declared silent when $\rho_j \geq 0.95$. The choice is not critical. Over $\tau \in [0.045, 0.3]$, counting accuracy on Libri2Mix stays within 96.1–98.2% on the development set and 97.9–98.9% on the test set (Figure 2). At test time, the rule uses the decoded streams alone, with neither references nor a counting module.

Results. Table 5 reports the results at $\tau = 0.1$. Counting accuracy is at least 96.43% at every speaker count. Errors are predominantly over-estimates, i.e., surplus streams declared active; under-estimates occur on only 0.07% of Libri3Mix and 0.53% of Libri4Mix mixtures. Since dWER is computed on the N streams matched to the references by PIT, it does not depend on τ. The difference between known and unknown speaker count therefore reflects the change of composition from $K = N$ to $K = 4$. This cost shrinks as N approaches K: it is 1.13 points on Libri2Mix, 0.13 points on Libri3Mix, and zero on Libri4Mix, where the two settings coincide. Given an upper bound K on the number of speakers, a single model can therefore separate mixtures of unknown speaker count without a dedicated counting module. Our evaluation covers $2 \leq N \leq K = 4$; single-speaker inputs are not evaluated and are also absent from training.

Table 6: Parameter count and MACs per module, normalized per second of 16 kHz audio. The separator is measured on the mixture alone, and the CFM for a single DiT forward pass; the cost at inference follows the scaling in the last column. Totals are given for SS with $S = 8$ sampling steps. The frozen WavLM that provides the training targets is not used at inference and is excluded.

Module	Params (M)	GMACs/s	Applied at inference
Separator	315.44	18.08	once; length KT (SS) or $T_e + T$ (TSE)
CFM (DiT), one forward pass	248.26	7.38	S steps $\times$ 2 (CFG) $\times$ K streams
Mel vocoder	113.73	5.69	once per stream
Total, SS (K=2)	677.43	285.5	
Total, SS (K=4)		578.2	

A.5 Computational Cost

Model size. Table 6 lists the parameter count and multiply–accumulate (MAC) cost of each module. The separator has exactly the parameter count of the WavLM-Large checkpoint from which it is initialized, 315.44 M. Input composition introduces no parameters, so both tasks are handled at no parameter cost over the pretrained model. The frozen WavLM that produces the training targets is a second instance of the same checkpoint and is not used at inference. The CFM and the Mel vocoder add 248.26 M and 113.73 M parameters, respectively, giving 677.43 M in total at inference.

Separator cost. The separator figure in Table 6, 18.08 GMACs per second of audio, corresponds to a single forward pass over the mixture alone. Input composition multiplies the sequence length by K in SS and extends it by the T_e enrollment frames in TSE. At the utterance lengths used here, the quadratic attention term accounts for only a minority of the separator's compute, so the cost grows close to linearly in K: relative to a single pass over the mixture, it is approximately $2.1\times$ for $K = 2$ and $4.6\times$ for $K = 4$.

Reconstruction cost. The CFM figure corresponds to one DiT forward pass. Generation takes S sampling steps, each requiring two forward passes under classifier-free guidance. Both the CFM and the vocoder are applied once per stream with shared parameters, so the reconstruction cost is linear in K and independent of the composition. For two-speaker SS with $S = 8$, reconstruction therefore requires approximately $(7.38 \times 8 \times 2 + 5.69) \times 2 \approx 248$ GMACs per second of mixture audio, against approximately $18.08 \times 2.1 \approx 38$ GMACs per second for the separator, for a total of about 286 GMACs/s. For TSE, a single stream is reconstructed, halving the reconstruction cost to approximately 124 GMACs/s, while the separator additionally processes the 3-s enrollment. In this configuration, inference cost is thus dominated by reconstruction rather than separation.

A.6 Reconstruction Ceilings

By construction, the separator determines which source each output contains: the CFM is trained to reconstruct the source paired with its conditioning representation, and no permutation search is involved in the reconstruction stage. This section therefore examines what the reconstruction path adds and where its limits lie. Table 7 compares two decoding paths: the L24 vocoder, which decodes layer-24 representations directly without observing the mixture, and the full back-end (CFM + Mel vocoder). The oracle setting replaces the separated representations with teacher representations of the clean references and thus gives the ceiling of each path; the Mel voc. row resynthesizes the clean reference Mel spectrograms and bounds the vocoder alone. The L24 and Mel vocoder oracles are identical across conditions, since neither path observes the mixture.

Speaker similarity. With oracle representations, the L24 vocoder reaches a SpkSim of only 0.65, indicating that layer-24 representations carry limited timbre information. The full back-end is subject to two successive ceilings: resynthesizing the clean reference Mel spectrograms with the Mel vocoder yields 0.93, and conditioning the CFM on oracle representations and the mixture yields 0.88 on Mix-Clean and 0.82 on Mix-Both, the latter being lower because timbre must then be recovered from a noisy mixture. On Mix-Clean, CoReSep reaches 0.86, so relative to perfect similarity, the losses due to the vocoder, the CFM, and separation are 0.07, 0.05, and 0.02, respectively. The

Table 7: Oracle ceilings and results with separated representations for the decoding paths on Libri2Mix. Oracle replaces the separated representations with teacher representations of the clean references; for Mel voc., it resynthesizes the clean reference Mel spectrograms. voc.: vocoder.

Decoder	Mix-Clean						Mix-Both					
	dWER (%)↓			SpkSim↑			dWER (%)↓			SpkSim↑		
	Oracle	TSE	SS	Oracle	TSE	SS	Oracle	TSE	SS	Oracle	TSE	SS
L24 voc.	0.76	1.51	1.29	0.65	0.58	0.57	0.76	3.92	3.58	0.65	0.54	0.52
Mel voc.	0.28	–	–	0.93	–	–	0.28	–	–	0.93	–	–
CFM + Mel voc.	0.48	1.22	0.85	0.88	0.86	0.86	0.60	3.93	3.65	0.82	0.77	0.76

vocoder ceiling alone equals the SpkSim of TF-GridNet (0.93) and lies just below that of MeanFlow-TSE (0.94); the gap to these baselines is thus bounded by the reconstruction back-end rather than by separation. On Mix-Both, the CFM ceiling (0.82) exceeds all baselines, whereas CoReSep reaches 0.77/0.76 for TSE/SS. Since the oracle differs from CoReSep only in the conditioning representation, this gap reflects imperfections of the separated representations under noise.

Content. The full back-end also yields lower dWER than the L24 vocoder on Mix-Clean, with both oracle (0.48% vs. 0.76%) and separated representations, whereas the two paths give nearly identical dWER on Mix-Both. We do not attribute this difference to a single cause. The lower oracle dWER suggests that part of it reflects output quality, for instance, the ASR model being less robust to L24-vocoded speech; the CFM may also partially compensate for imperfect separated representations using the mixture.