Blind Source Separation for Fetal PPG with Rate-Based Proxy Supervision

Weitai Qian[1] **Rishad Joarder**[1] **Randall Fowler**[1] **Soheil Ghiasi**[1]
[1]Department of Electrical and Computer Engineering, UC Davis
{wtqian, rrjoarder, rlfowler, ghiasi}@ucdavis.edu

Abstract

Separating fetal photoplethysmography (PPG) from non-invasively acquired transabdominal recordings is a critical step toward reliable estimation of fetal arterial oxygen saturation ($fSpO_2$). We present an end-to-end attention-based framework that extracts fetal components from mixed maternal and fetal signals. Since ground-truth fetal PPG cannot be collected in practice, the model is pre-trained on physics-informed synthetic mixtures and evaluated on *in-vivo* data from pregnant ovine studies. To enable deployment in real-world settings, we employ a multi-objective design in which separation is paired with rate estimation, serving as a proxy for assessing separation quality when ground truth is unavailable. On *in-vivo* recordings, the proposed framework improves downstream $fSpO_2$ estimation, reducing mean absolute error by 33.3%, lowering the standard deviation of error by 30.7%, and increasing correlation with reference values by 26.9% compared to baseline methods. Overall, this end-to-end pipeline has the potential to enhance real-time $fSpO_2$ estimation and advance non-invasive fetal monitoring in clinical practice.

1 Introduction

Transabdominal photoplethysmography (PPG) records pulsatile optical signals through the maternal abdomen and forms the sensing foundation of Transabdominal Fetal Oximetry (TFO). Because the optical path traverses multiple heterogeneous tissue layers (abdominal, adipose, uterine, and fetal), the acquired signal is a complex mixture of maternal and fetal PPG, respiration, and additional physiological or environmental noise. Accurate estimation of fetal oxygen saturation ($fSpO_2$) therefore requires reliable separation of the fetal PPG component from this mixture.

This problem is an instance of *Blind Source Separation* (BSS), a longstanding challenge in signal processing where the goal is to recover constituent sources from their mixtures without explicit knowledge of mixing dynamics. In recent years, deep learning has transformed BSS, particularly in audio processing. Time-domain models such as TasNet and its variants demonstrated that learning directly from raw waveforms with an encoder–decoder architecture can outperform traditional time–frequency methods, enabling true end-to-end optimization [10, 11, 14]. Dual-path architectures and attention-based extensions further improved the modeling of long-range temporal dependencies [12, 15, 19]. Inspired by these advances, similar strategies have been successfully applied in biomedical domains, including EEG artifact removal, fetal ECG separation, and bioacoustic signal separation [1–3, 5, 8].

In this work, we present a supervised end-to-end framework for transabdominal fetal PPG separation. Unlike audio BSS, where clean source signals are typically available for supervision, direct supervision is infeasible here since ground-truth fetal PPG cannot be collected in practice. To address this, we pre-train using physics-informed synthetic mixtures and evaluate on *in-vivo* recordings from animal studies. Our model employs a multi-objective formulation in which separation is paired with

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Workshop on Learning from Time-Series for Health.

an auxiliary prediction task that serves as a proxy for separation quality in real-world deployments. This design improves interpretability and supports practical use in safety-critical fetal monitoring applications, even in the absence of ground-truth fetal signals.

2 Methods

We model the observed mixed-PPG signal $x(t)$ as the additive combination of C source signals $s_1(t), s_2(t), \ldots, s_C(t) \in \mathbb{R}^{1\times T}$, where T denotes the signal duration. In this work, we consider $C = 3$ sources corresponding to maternal PPG, fetal PPG, and maternal respiration. Each source is represented as a quasi-periodic waveform, where individual cycles are modeled by a function $g_i(\cdot)$ and modulated by time-varying amplitude $a_i(t)$ and instantaneous frequency $\omega_i(t)$, i.e., $s_i(t) = a_i(t) \cdot g_i(\omega_i(t))$. The goal is to recover $\{s_i(t)\}_{i=1}^{C}$ from $x(t)$ with high fidelity, ensuring that the reconstructed fetal component retains information critical for downstream $fSpO_2$ estimation.

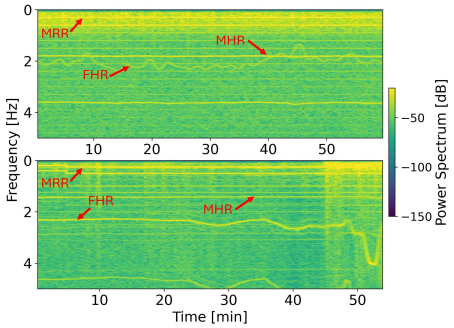

Figure 1: **Spectrograms of mixed PPG.** Top: synthesized mixture. Bottom: *in-vivo* sheep experiment with labeled Maternal Heart Rate (MHR), Fetal Heart Rate (FHR), and Maternal Respiration Rate (MRR).

Dual-dataset setup. To supervise model training, we construct a synthetic dataset by combining three components: (i) quasi-periodic waveform generation for $g_i(\cdot)$ using a time-domain PPG generator [20], (ii) physics-compliant Monte Carlo simulations to model source magnitudes $a_i(t)$ [13], and (iii) source frequency profiles derived from the CTU-CHB and MIMIC-III datasets [4, 7]. This process yields approximately 2,040 hours of mixed-PPG recordings, which are used for training and validation of the separation network. To evaluate performance in realistic settings, we employ an *in-vivo* dataset collected during controlled fetal hypoxia experiments in pregnant ewe models. This dataset consists of roughly 8 hours of mixed-PPG signals without ground-truth source time-series. Separation performance on the *in-vivo* recordings is therefore assessed indirectly through a downstream $fSpO_2$ estimation task.

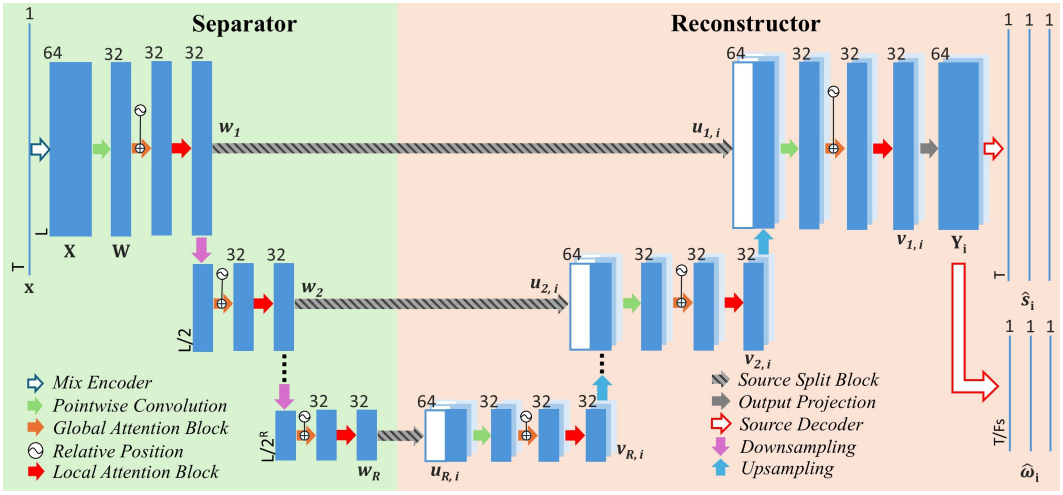

Figure 2: **Separation-Reconstruction network architecture.** The model consists of a multi-stage separator followed by a reconstructor. Each stage contains a source-split block with an early-split mechanism that partitions the latent mixture representation into C source components. The network produces both reconstructed source signals and their corresponding frequency series.

Preprocessing and Data Augmentation. Each mixed PPG recording is normalized to zero mean and unit variance, with each source $s_i(t)$ scaled by the variance of the mixture $x(t)$ and re-centered to preserve linear additivity $x(t) = \sum_{i=1}^{C} s_i(t)$ while standardizing inputs for training. The mixture

is then segmented into fixed-length windows $x \in \mathbb{R}^{1\times T}$, ensuring compatibility with the network input and enabling streaming or real-time separation; we set T to 5 minutes in this work.

To increase variability and mitigate overfitting on the synthetic dataset, we apply stochastic data augmentation. Given a mixture x, an augmented sample $\tilde{x}$ is generated by randomly applying one or more of the following transformations:

- Additive white noise: $\tilde{s}_i(t) = s_i(t) + \epsilon(t),\ \epsilon(t) \sim \mathcal{N}(0, \sigma^2)$.
- Time reversal of n sources, $0 < n < C$: $\tilde{s}_i(t) = s_i(T - t)$.
- Amplitude scaling: $\tilde{s}_i(t) = \alpha_i s_i(t),\ \alpha_i \sim \mathcal{U}(\alpha_{\min}, \alpha_{\max})$.
- Random low-pass filtering with cutoff $f_c \sim \mathcal{U}[10, 20]$ Hz.
- Temporal masking: $\tilde{s}_i(t) = 0,\ t \in [t_0, t_0 + T'],\ 0 < T' < T$.

These augmentations improve robustness by exposing the model to signal variability, artifacts, and distortions commonly observed in *in-vivo* recordings, thereby enhancing generalization beyond the synthetic training set.

Separation–Reconstruction Network. The proposed network follows a separator–reconstructor architecture (Fig. 2) inspired by U-Net and related variants [16–18]. It performs source decomposition and reconstruction through a hierarchical refinement process over R stages, where the separation path progressively downsamples the input and the reconstruction path symmetrically upsamples it.

The input mixture $\mathbf{x} \in \mathbb{R}^{1\times T}$ is first encoded into a latent representation $\mathbf{X} \in \mathbb{R}^{F\times L}$ using strided convolution, where F is the number of filters and L is the compressed sequence length. In this work, we set $F = 64$ to balance representational capacity and efficiency. A pointwise convolution reduces this to $\mathbf{W} \in \mathbb{R}^{F/2\times L}$ to further improve efficiency. The downsampling path then processes $\mathbf{W}$ across R stages, where at stage j ($j = 1, \ldots, R$) the temporal resolution is reduced to L/j, enabling multi-scale feature extraction.

In the reconstruction path, source-specific features $\mathbf{u_{j,i}}$ from stage j are fused with the upsampled features $\mathbf{v_{j+1,i}}$ from the next stage to form $\mathbf{v_{j,i}}$, supporting hierarchical reconstruction. At the final stage, $\mathbf{v_{1,i}} \in \mathbb{R}^{F/2\times L\times C}$ is projected to $\mathbf{Y_i} \in \mathbb{R}^{F\times L\times C}$ and transformed back to the time domain via transposed 1D convolution, producing the separated signals $\hat{s}_i \in \mathbb{R}^{1\times T}$ for $i = 1, \ldots, C$.

Multi-Task Learning with Proxy Supervision. In *in-vivo* monitoring, ground-truth source signals are inaccessible, making direct evaluation of separation performance infeasible. To address this, we introduce an auxiliary objective that predicts the frequency series (e.g., heart and respiration rates) of each source, serving as a proxy for separation quality and providing a real-time reliability measure for downstream use.

The network is trained with two objectives: (i) signal reconstruction and (ii) frequency estimation, both derived from a shared latent representation. One decoder outputs reconstructed sources $\hat{s}_i$, while another predicts source frequencies $\hat{\omega}_i$ (Fig. 2). The reconstruction loss is defined by negative SDR, and the frequency estimation loss by MSE:

$$L_s = -10 \cdot \log_{10}\left(\frac{1}{C}\sum_{i=1}^{C}\frac{\|s_i\|^2}{\|s_i - \hat{s}_i\|^2}\right), \quad L_\omega = \frac{1}{C}\sum_{i=1}^{C}\|\omega_i - \hat{\omega}_i\|^2. \tag{1}$$

Since these objectives operate in different unit spaces, we use *Dynamic Weight Averaging* (DWA) [9] to balance their contributions during training. Finally, to address permutation ambiguity in multi-source separation, we adopt *Permutation Invariant Training* (PIT) [21], minimizing the joint loss $L = \{L_s, L_\omega\}$ over all possible pairings between predictions and targets.

3 Results

Signal Reconstruction on Synthetic Time-Series

Our Separation–Reconstruction network was constrained to under 1 MB for real-time deployment, with baseline models trained under the same limit. Performance was evaluated on 408 held-out synthesized recordings using a rolling 5-minute SDR, providing a more robust measure than full-sequence SDR by reducing bias from transient high-energy segments.

Table 1: **Performance comparison on the synthesized validation dataset.** SDR is computed with a 5-minute rolling window. All models are restricted to under 1 MB for fair comparison.

Method	SDR($s_i, \hat{s_i}$) (dB)	SDR($s_f, \hat{s_f}$) (dB)
EMD	1.23	0.76
Wave-U-Net	10.1	8.21
Conv-TasNet	10.35	8.87
Sepformer	16.9	15.4
This work	**18.27**	**16.2**

Table 1 reports SDR($s_i, \hat{s}_i$) for all sources and SDR($s_f, \hat{s}_f$) specifically for fetal PPG. Our method consistently outperforms both the classical baseline (EMD [6]) and deep learning models (Wave-U-Net [18], Conv-TasNet [11], Sepformer [19]). Notably, it achieves the highest SDR on fetal PPG, the most clinically relevant component.

Estimating fSpO$_2$ with Separated Fetal PPG on in-vivo Dataset

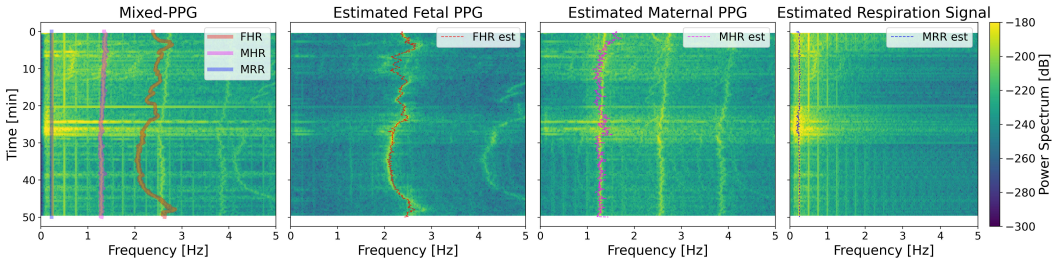

Figure 3: **Example separation of in-vivo mixed PPG.** A 50-minute recording is separated into fetal PPG, maternal PPG, and respiration. The left spectrogram shows the mixture with labeled FHR, MHR, and MRR, while the subsequent plots display the reconstructed sources and their estimated rates.

After training on synthesized mixed-PPG, the separation–reconstruction network is directly applied to in-vivo data without fine-tuning. Figure 3 shows an example from a 50-minute hypoxia experiment in a pregnant sheep model, where fetal, maternal, and respiratory signals are separated with their corresponding rate estimates.

Table 2: **fSpO$_2$ estimation on in-vivo data.** Comparison with baseline lock-in detection [13].

Metric	fSpO$_2$ estimation methods		
	Baseline	**this work**	Impr.
MAE	7.25	**4.83**	33.3%
σ(MAE)	6.06	**4.2**	30.7%
Correlation	0.67	**0.85**	26.9%

To assess the utility of separation, we evaluate a downstream task: fSpO$_2$ estimation in the TFO setting. Light intensity ratios r are computed from the AC/DC components of the fetal PPG. The baseline method derives r using lock-in detection, whereas this work extracts it from the separated fetal signal via envelope detection. For a fair comparison, both methods use the same sensor-fusion MLP. To incorporate separation reliability, we provide the MLP with a proxy for separation quality—scaled MAE between estimated and ground-truth FHR ($\hat{\omega}_f$ vs. ω_f) in this work.

As summarized in Table 2, using separated fetal PPG with a separation quality proxy improves downstream performance. This work reduces MAE by 33.3%, decreases error variance by 30.7%, and raises correlation with ground truth by 26.9%. These results highlight the advantage of source-aware signal quality assessment for robust fSpO$_2$ estimation in real-world monitoring.

4 Conclusion

This work presents an end-to-end pipeline for fetal PPG separation, trained on physics-compliant synthetic data and evaluated on in-vivo recordings. Incorporating a separation-quality proxy improves fSpO$_2$ estimation over baselines, underscoring the potential of source-aware separation for robust, noninvasive fetal monitoring in real-time settings.

References

[1] Peter C Bermant. Biocppnet: automatic bioacoustic source separation with deep neural networks. *Scientific Reports*, 11(1):23502, 2021.

[2] Michael Chan, Venu G Ganti, and Omer T Inan. Respiratory rate estimation using u-net-based cascaded framework from electrocardiogram and seismocardiogram signals. *IEEE journal of biomedical and health informatics*, 26(6):2481–2492, 2022.

[3] Chun-Hsiang Chuang, Kong-Yi Chang, Chih-Sheng Huang, and Tzyy-Ping Jung. Ic-u-net: a u-net-based denoising autoencoder using mixtures of independent components for automatic eeg artifact removal. *NeuroImage*, 263:119586, 2022.

[4] Václav Chudáček, Jiří Spilka, Miroslav Burša, Petr Janků, Lukáš Hruban, Michal Huptych, and Lenka Lhotská. Open access intrapartum CTG database. *BMC Pregnancy Childbirth*, 14:16, January 2014.

[5] Lei Hu, Wenjie Cai, Ziyang Chen, and Mingjie Wang. A lightweight u-net model for denoising and noise localization of ecg signals. *Biomedical Signal Processing and Control*, 88:105504, 2024.

[6] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, and H. H. Liu. The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 454(1971):903–995, March 1998. doi: 10.1098/rspa.1998.0193.

[7] Alistair E W Johnson, Tom J Pollard, Lu Shen, Li-Wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Sci. Data*, 3(1):160035, May 2016.

[8] Kwang Jin Lee and Boreom Lee. End-to-end deep learning architecture for separating maternal and fetal ecgs using w-net. *IEEE Access*, 10:39782–39788, 2022. doi: 10.1109/ACCESS.2022.3166925.

[9] Shikun Liu, Edward Johns, and Andrew J Davison. End-to-end multi-task learning with attention. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1871–1880, 2019.

[10] Yi Luo and Nima Mesgarani. Tasnet: time-domain audio separation network for real-time, single-channel speech separation. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 696–700. IEEE, 2018.

[11] 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.

[12] Yi Luo, Zhuo Chen, and Takuya Yoshioka. Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 46–50. IEEE, 2020.

[13] Weitai Qian, Rishad Raiyan Joarder, Randall Fowler, Begum Kasap, Mahya Saffarpour, Kourosh Vali, Tailai Lihe, Aijun Wang, Diana Farmer, and Soheil Ghiasi. Transabdominal fetal oximetry via diffuse optics: Principled analysis and demonstration in pregnant ovine models, 2025. URL `https://arxiv.org/abs/2509.21594`.

[14] William Ravenscroft, Stefan Goetze, and Thomas Hain. Att-tasnet: Attending to encodings in time-domain audio speech separation of noisy, reverberant speech mixtures. *Frontiers in Signal Processing*, 2:856968, 2022.

[15] Joel Rixen and Matthias Renz. Qdpn-quasi-dual-path network for single-channel speech separation. In *Interspeech*, pages 5353–5357, 2022.

[16] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. *CoRR*, abs/1505.04597, 2015. URL `http://arxiv.org/abs/1505.04597`.

[17] Ui-Hyeop Shin, Sangyoun Lee, Taehan Kim, and Hyung-Min Park. Separate and reconstruct: Asymmetric encoder-decoder for speech separation. In *2024 Conference on Neural Information Processing Systems*, 2025. URL `https://arxiv.org/abs/2406.05983`.

[18] Daniel Stoller, Sebastian Ewert, and Simon Dixon. Wave-u-net: A multi-scale neural network for end-to-end audio source separation. *arXiv preprint arXiv:1806.03185*, 2018.

[19] 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)*, pages 21–25. IEEE, 2021.

[20] Qunfeng Tang, Zhencheng Chen, Rabab Ward, and Mohamed Elgendi. Synthetic photoplethysmogram generation using two gaussian functions. *Scientific Reports*, 10(1):13883, Aug 2020. ISSN 2045-2322. doi: 10.1038/s41598-020-69076-x. URL `https://doi.org/10.1038/s41598-020-69076-x`.

[21] 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)*, pages 241–245. IEEE, 2017.

A Technical Appendices and Supplementary Material

Physics-Compliant Monte-Carlo Simulation for Time-Series Synthesis

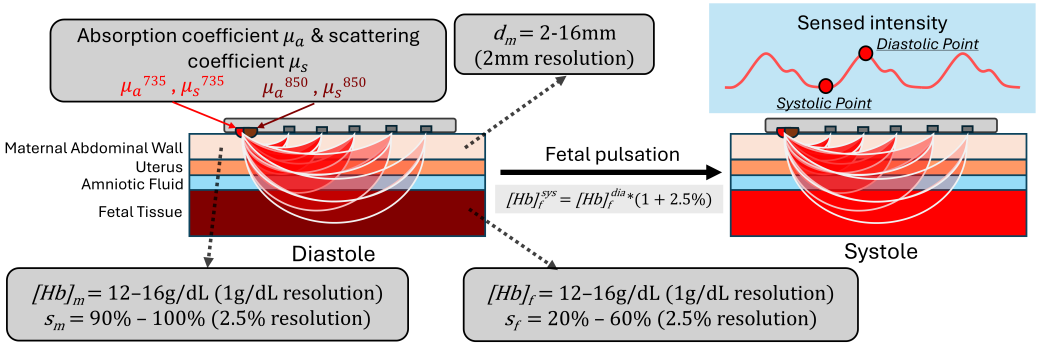

Figure 4: **Multi-layer tissue model for MC simulation.** It includes one geometric parameter (d_m) and four hemodynamic parameters: $[Hb]_m$, s_m, $[Hb]_f$, and s_f. [13]

A modified MCXtreme simulator modeled photon transport in a four-layer maternal–fetal tissue stack [13]. Sequential 735- and 850-nm pencil beams illuminated the model, with reflected light measured by 20 detectors (10–95 mm SDD). Simulations varied maternal wall thickness (d_m), maternal and fetal hemoglobin concentration and oxygenation ($[Hb]_m$, s_m, $[Hb]_f$, s_f). Fetal pulsation was represented by two static states generated via a 2.5% perturbation in $[Hb]_f$. Over two million combinations across parameters, wavelengths, and pulsation states formed the synthetic dataset.

Early Split Block

The early-split strategy [17] introduces source-specific branches early in the network, unlike methods that separate only at the output [11, 12, 18]. At stage j, latent features $\mathbf{w}_j$ are expanded via pointwise convolutions with GLUs and normalization to produce $\mathbf{u}_j$, enabling early, source-aware refinement. Skip connections between separation and reconstruction preserve detail.

Global and Local Attention Blocks

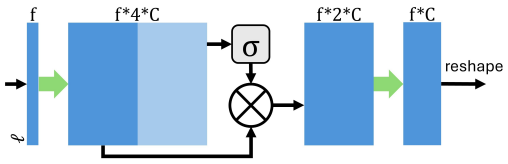

Figure 5: **Source split block.** This block is deployed at each stage to perform early splitting of sources within the latent space, enabling source-specific feature extraction.

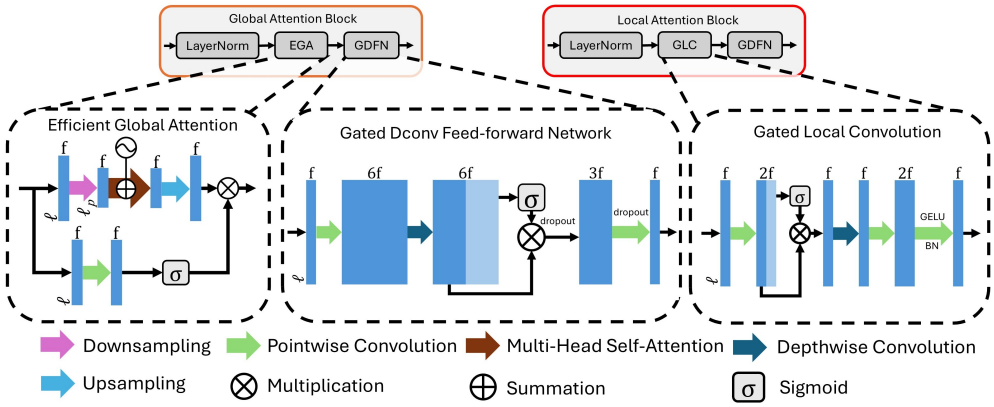

Figure 6: **Sub-modules of GAB and LAB.** Both blocks share a similar structure, with GAB using global attention and LAB using local convolution to capture complementary long- and short-range features.

Each stage uses cascaded Global (GAB) and Local Attention Blocks (LAB). Both apply layer normalization and a Gated Depthwise Feed-forward Network (GDFN). GAB uses downsampled self-attention (EGA) for long-range context, while LAB uses Gated Local Convolutions for fine-grained features. Together they provide complementary global–local modeling for hierarchical separation.