Supplementary Material

Supplement to Section 5: Training Configuration

Unless otherwise noted, all spatiotemporal models are trained for 15 epochs with a fixed learning rate of 1×10^{-3} and a mini-batch size of 4. We fixed random seeds and enforced deterministic PyTorch settings for reproducibility. For validation, testing, and inference, we use a batch size of 8 to better utilize GPU memory.

We did not employ early stopping; instead, for each experiment we select the checkpoint with the best validation performance according to the subject-averaged $\mathcal{L}_{\text{PSD}}$ in (10) among the 15 epochs. In practice, the validation loss stabilizes well before the final epoch, and we did not observe divergence across runs.

Our codebase also supports alternative training objectives, including negative Pearson correlation and time-domain mean squared error, but unless otherwise specified, all results in Section 5 use the PSD-MSE objective. For reproducibility, the exact hyperparameters and loss configurations for each experiment are specified in the YAML configuration files included in our public code release.

Supplement to Section 5.1: RAFT Optical Flow Failure

While classical optical flow methods (Coarse2Fine, DeepFlow, Farnebäck, PCAFlow, and TV-L1) yield consistent performance across models, we observe a dramatic degradation when using RAFT. Across all spatiotemporal networks, RAFT produces up to $5\times$ higher MAE and RMSE, and even negative correlations, indicating a systematic failure rather than stochastic variance. As visualized in Figure 5, RAFT predicts dense, large-magnitude motion across nearly the entire frame, including static background regions, instead of isolating subtle, periodic chest motion.

We hypothesize that this behavior reflects a domain mismatch. RAFT was developed and evaluated primarily on large-displacement, high-texture motion benchmarks such as Sintel and KITTI [32], whereas our infrared infant videos exhibit low texture, low signal-to-noise ratio (SNR), and only minute thoracoabdominal motion. In this setting, RAFT's dense all-pairs correlation volume may overinterpret small fluctuations and sensor noise as coherent motion, leading to physiologically implausible flow fields.

By contrast, classical variational and coarse-to-fine methods such as TV-L1 explicitly regularize the flow with spatial smoothness terms and robust data penalties, and enforce multi-scale consistency through pyramid-based warping [31, 40]. These priors are known to suppress high-frequency noise and favor piecewise-smooth motion, and have been successfully applied to weak-texture, low-contrast medical imagery [1], consistent with our observation that they produce flows more localized to the thoracoabdominal region and support substantially lower respiration estimation errors.

Supplement to Section 5.3: Chest ROI Inconsistency

We additionally investigate why the more concentrated chest ROI does not consistently outperform the coarser body ROI. Our chest ROI is defined as a square inscribed within the body box (using its shorter side) and shifted slightly toward the head. As illustrated in Figure 4, this crop emphasizes the upper torso but truncates part of the lower abdomen, where supine infant breathing motion is often pronounced. Meanwhile, head, shoulder, and arm movements (as well as blankets or toys) are more prevalent in the upper region and therefore receive greater weight inside the chest ROI.

By contrast, the body ROI preserves the full thoracoabdominal motion field while still restricting the input to the infant torso, giving the spatiotemporal networks sufficient context to internally attend to the most informative subregions. These factors provide a plausible explanation for why the body ROI yields more reliable improvements, whereas the chest ROI produces mixed results across architectures.

6.1. Supplement to Section 5.4: Reproducibility in Manne et al. [23]

Regrettably, we could not recall or uncover documentation of the specific model and training configuration used in Manne et al., nor could we obtain precisely the same metric results with reconstructions, despite extensive efforts on both fronts.

Our best attempt to reproduce the AIRFlowNet configuration used in [23] yielded a model achieving a MAE of 3.84 BPM, under what we believe to be the original train–test split, with three subjects chosen for training and five for testing. We then evaluated this replicate model under *all* $\binom{8}{3} \times 3 = 168$ train–test splits with three subjects chosen for training (one held for validation), and plotted the results in Figure 8. These reveal a high variability in MAE by split choice (consistent with our findings on fold-variability in Figure 6), and a particularly low MAE of 3.84 BPM achieved by the split likely used in Manne et al., compared to mean MAEs of 6.16 BPM by split. We believe these higher MAEs in the 5.5–6.5 BPM range more accurately reflect the true performance Manne et al.'s model in the AIR-125 dataset, and that our present results in the 3.7–4.0 BPM range on AIR-400 reflect the current, generalizable state-of-the-art performance of infant respiration rate estimation from spatiotemporal models.