Breathing Measurement From Video Using Dense Point Tracking

Sofia Taouhid
MSc in Data Science
EPFL
Lausanne, Switzerland

Supervisors: Mallory Wittwer, Fariza Sabit, Dr. Edward Ando, Prof. Dolaana Khovalyg
Center for Imaging and ICE Laboratory
EPFL

***Abstract*—Non-contact respiration monitoring is an attractive alternative to contact sensors in clinical and tele-health settings. This report evaluates whether dense, long-term point tracking can recover respiratory motion reliably across heterogeneous cameras (RGB, grayscale, thermal). We propose an end-to-end pipeline that crops a torso region of interest (chest, abdomen or person segmentation), applies local contrast enhancement, tracks a dense grid of points using CoTracker, and finally aggregates trajectories into a 1D breathing waveform using either a robust median displacement or a PCA-based dominant motion component. From this waveform, we estimate a single respiratory rate per phase interval using Welch spectral analysis and extract a breath-amplitude proxy (peak-to-valley) and explore a linear calibration to tidal volume. Evaluation is performed against synchronous COSMED measurements of respiratory rate Rf, tidal volume VT, and minute ventilation VE. In addition to rate error, we report waveform-level agreement metrics (correlation, lag, and magnitude-squared coherence) between the video-derived waveform and reference signals.**

I. Introduction

Non-contact respiration monitoring aims to estimate breathing signals from video without attaching sensors to the subject. This is valuable in situations where contact-based systems can be uncomfortable, restrictive, or impractical for long-term recording. From a computer vision perspective, breathing manifests as a subtle but structured motion: during inhalation and exhalation, the chest and abdomen produce a quasi-periodic displacement that can be captured by cameras given sufficient resolution and stability.

A major challenge in this domain is the low signal-to-noise ratio; the amplitude of respiratory motion is often small compared to other sources of variation, such as posture changes, head movements, and tracking drift. This necessitates the use of dense, robust motion estimation methods. Recent advances in point tracking have introduced transformer-based models, such as CoTracker [1], which jointly track distinct points across video frames. CoTracker utilizes an attention mechanism to handle occlusions and complex motions, making it a strong candidate for extracting subtle physiological signals.

In this work, we investigate three primary questions:

1) Can dense point tracking produce stable breathing waveforms and accurate respiratory-rate estimates?
2) Which anatomical region is most informative for motion extraction (chest vs. abdomen vs. segmentation)?
3) Can multi-view fusion improve robustness of breathing measurement?

II. Related Work

Vision-based respiration monitoring has been studied for over a decade as a promising alternative to contact sensors in clinical, home, and sleep settings. Most approaches fall into two broad families: (i) *appearance-based* methods that recover a physiological waveform from subtle intensity/color changes (often inspired by remote photoplethysmography pipelines), and (ii) *motion-based* methods that measure thoracoabdominal displacement or deformation over time. The latter is particularly relevant for respiration, where chest-wall motion provides a direct mechanical correlate of breathing and can be robust even when color cues are weak or lighting varies.

Classical motion extraction and signal enhancement

Early motion-based pipelines typically use frame differencing, block matching, or optical-flow-like measurements inside a manually or automatically defined torso ROI, followed by band-limited filtering and spectral peak selection to estimate respiratory rate. While these methods can work well in controlled scenes, they often degrade over long sequences due to drift, low texture (e.g., plain clothing), and confounds such as posture changes. A related line of work uses motion amplification to make periodic motion visually salient. Eulerian Video Magnification [2] and subsequent phase-based formulations [3] can reveal subtle periodic signals, including breathing-related motion, but they are not always reliable for quantitative estimation: amplification can also boost noise and non-respiratory motion, and the output is sensitive to parameter choices (spatial bands, temporal pass-bands, magnification factor).

Optical flow, feature tracking, and learning-based correspondence

Feature tracking provides an alternative to dense flow by following sparse keypoints through time. Classical trackers

(e.g., Lucas–Kanade) estimate motion by local image registration [4] and can be efficient, but they are prone to losing points under low texture, specularities, or partial occlusions. Dense optical flow models improve robustness by enforcing global consistency; modern learning-based flow such as RAFT [5] provides strong performance on generic motion benchmarks, yet flow fields can still exhibit low-frequency drift and may be unstable in homogeneous regions—precisely the regime encountered in subtle physiological motion. These limitations motivate the use of robust aggregation (median-type statistics) and careful ROI selection to suppress outliers and background leakage.

Region-of-interest strategies

Because respiration is localized and low-amplitude, most systems rely on ROI selection to raise signal-to-noise ratio. ROIs can be manual (fixed bounding boxes) or automatic via face/body detection, keypoints, or segmentation masks. Massaroni et al. [6] emphasize that ROI choice strongly affects performance and show that even small, anatomically meaningful regions can yield accurate respiratory signals when the extracted waveform is appropriately filtered and analyzed.

Long-term dense point tracking and multi-camera robustness

Recent transformer-based correspondence models enable long-term point tracking by reasoning jointly over space and time. CoTracker [1] produces dense trajectories with explicit visibility estimates, which is well-suited to respiration where (i) motion is subtle but coherent over the torso and (ii) some points may temporarily fail due to occlusions or low contrast. Our work builds on this capability and contributes a respiration-specific pipeline around dense trajectories: local contrast normalization (CLAHE), robust waveform construction (median displacement or PCA dominant motion), and a waveform-level evaluation against synchronous COSMED signals (correlation/lag/coherence), not only RR error. Finally, while multi-camera geometry is commonly used in 3D reconstruction, it is less frequently integrated into respiration pipelines; we therefore compute and validate multi-view calibration to enable principled future 3D fusion and cross-view consistency checks, and we empirically study whether simple multi-view aggregation (camera-median fusion) improves waveform fidelity across heterogeneous sensors.

III. Experimental Setup

To ensure geometric diversity and signal robustness, we designed a custom acquisition environment. We defined acquisition geometry, device selection, breathing protocols, and synchronization procedure.

Breathing protocol

Participants performed a single continuous recording consisting of a sequence of instructed respiratory phases (two tidal, one fast and one deep breathing) over a period of 30 to 50s, separated by short rest intervals. These phases differ primarily in respiratory frequency (Rf) and amplitude-related measures (VT), but can also potentially introduce non-respiratory motion such as minor posture drift.

TABLE I
Camera Specifications and Settings

ID	Device	Modality	Res.	FPS
Cam0	Olympus	RGB	1920×1080	25
Cam1	iPhone 16 Pro	RGB	1080×1920	30
Cam2	FLIR	Thermal/RGB	640×480	15
Cam3	Allied Vision	Grayscale	4024×3036	9

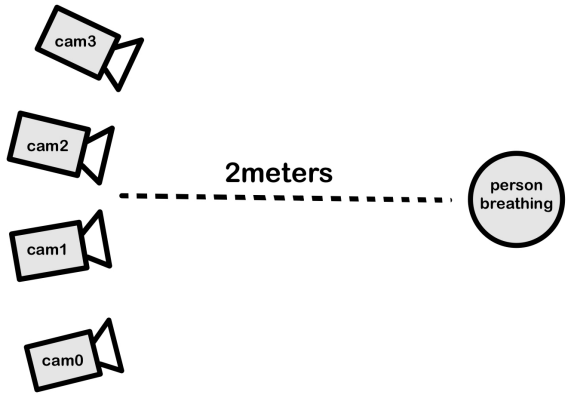


Fig. 1. Experimental setup and camera placement.

Acquisition and Synchronization

All cameras were fixed and positioned approximately $2.0\,\mathrm{m}$ from the subject (Fig. 1), oriented to ensure the upper body was clearly visible (Fig. 2). Early trials revealed that silhouette detection is prone to failure when the head is excluded from the frame. Consequently, the final framing included both the head and the upper torso to facilitate robust Region of Interest (ROI) detection. At the end, for each participant, we had one continuous sequence per camera that we segmented into four phase intervals using synchronization markers. This yields 2 participants $\times$ 4 phase segments $\times$ 4 cameras $=$ 32 phase-specific video segments used for our analysis.
Phase boundaries were defined once per participant and applied consistently across cameras using the synchronized time base.

Although devices record at different native FPS (Table I), for comparability and computational consistency, all videos are processed at a synchronized frame rate. We chose the lowest frame rate across the four cameras ($f_s \approx 9\,\mathrm{fps}$) and downsampled all videos to that frame rate using linear interpolation.

Ground Truth Signals (COSMED)

Among the information provided by COSMED, we obtained time-stamped respiratory rate $Rf(t)$ (breaths/min), tidal volume $VT(t)$ (L), and minute ventilation $VE(t)$ (L/min). While recording our data, we used synchronization markers to segment each phase and make it linkable to the corresponding video when processing the data. For each phase interval,

Fig. 2. Frame example.

we extracted the corresponding COSMED segment, converted timestamps to seconds relative to the phase starting point, and resampled to a uniform time grid with $\Delta t = 0.25\,\mathrm{s}$. This produces continuous reference waveforms $VT_{\mathrm{ref}}(t)$ and $VE_{\mathrm{ref}}(t)$, as well as a reference mean respiratory rate computed from $Rf(t)$ over the analyzed interval. We chose the mean Rf over an interval instead of Rf over time because of our short intervals and the fact that the breathing is in general constant on the same interval.

Multi-view Calibration Acquisition

In addition to the breathing recordings, we captured calibration images to estimate multi-view camera geometry (intrinsics and extrinsics) for future 3D breathing analysis and cross-view fusion. A planar checkerboard target was recorded from multiple viewpoints, ensuring strong pose diversity (rotations and translations) and broad image coverage for each camera.

To form usable multi-view sets, calibration frames were synchronized across cameras by filename (common basenames), and we retained only images where checkerboard corners were successfully detected in all four views.

Multi-view Calibration Estimation

Let $\mathbf{X}_j \in \mathbb{R}^3$ denote the 3D coordinates of checkerboard corner j in the checkerboard coordinate system, with known square size. For each camera c, we estimate intrinsic parameters $(\mathbf{K}_c, \mathbf{d}_c)$ (pinhole matrix and distortion coefficients) using standard monocular calibration [7]:

$$\mathbf{x}_{c,j} \sim \pi(\mathbf{K}_c, \mathbf{d}_c;\, \mathbf{R}_{cB}, \mathbf{t}_{cB};\, \mathbf{X}_j),$$

where $\pi(\cdot)$ denotes projection with distortion, and $(\mathbf{R}_{cB}, \mathbf{t}_{cB})$ is the pose of the checkerboard in camera coordinates for each calibration image.

We select a reference camera (Cam0) and estimate extrinsics for each camera c relative to Cam0, yielding rigid transforms $(\mathbf{R}_{0\to c}, \mathbf{t}_{0\to c})$. Extrinsics are solved by stereo calibration between Cam0 and each other camera while keeping intrinsics fixed, returning a single relative transform per camera.

Calibration Validation via Cross-view Reprojection

Calibration quality is verified using a cross-view reprojection check. For a given synchronized calibration image, we estimate the checkerboard pose in the *source* camera (Cam0) by solving PnP, yielding $(\mathbf{R}_{0B}, \mathbf{t}_{0B})$. The corresponding 3D corner coordinates in the Cam0 frame are

$$\mathbf{X}_{0,j} = \mathbf{R}_{0B}\mathbf{X}_j + \mathbf{t}_{0B}.$$

Using the Cam0-referenced extrinsics, we compute the rigid transform from Cam0 to a *target* camera. Since Cam0 is the source view, the transform is simply

$$\mathbf{R}_{\mathrm{tgt}\leftarrow 0} = \mathbf{R}_{0\to\mathrm{tgt}}, \qquad \mathbf{t}_{\mathrm{tgt}\leftarrow 0} = \mathbf{t}_{0\to\mathrm{tgt}}.$$

Accordingly, the checkerboard pose estimated in Cam0 is transferred to the target camera as

$$\mathbf{R}_{\mathrm{tgt}B} = \mathbf{R}_{0\to\mathrm{tgt}}\,\mathbf{R}_{0B}, \qquad \mathbf{t}_{\mathrm{tgt}B} = \mathbf{R}_{0\to\mathrm{tgt}}\,\mathbf{t}_{0B} + \mathbf{t}_{0\to\mathrm{tgt}}.$$

Finally, we reproject all checkerboard corners into the target image using $(\mathbf{K}_{\mathrm{tgt}}, \mathbf{d}_{\mathrm{tgt}})$ and compare them to the detected target corners. Low pixel reprojection error indicates consistent multi-view geometry. Figure 3 shows a qualitative example (detected corners vs. reprojected corners).

IV. Methods

Overview

Given a phase interval, our pipeline produces:

- a video-derived breathing waveform $s(t)$ from dense trajectories,
- a single respiratory-rate estimate $\widehat{Rf}$ from the waveform spectrum,
- breath-level amplitude events $\{(t_k, a_k)\}$ used as a proxy for tidal volume and used in an exploratory linear calibration to VT.

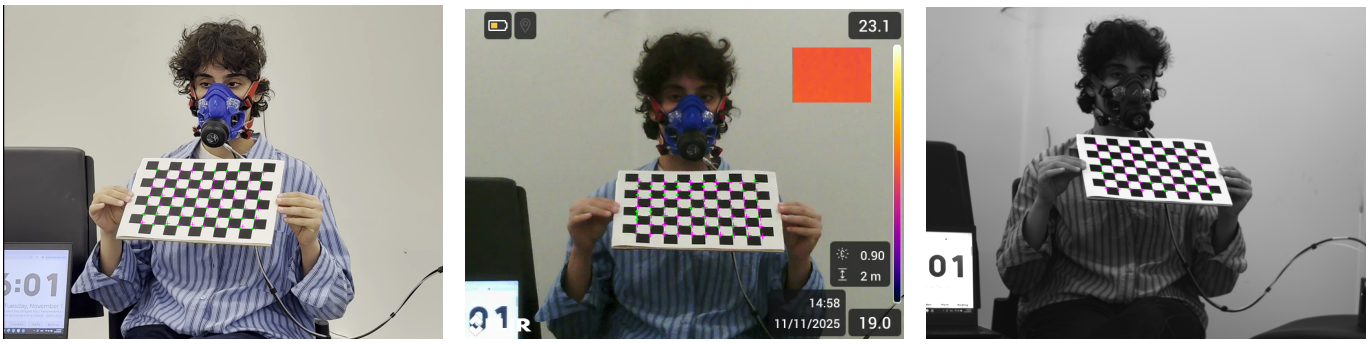
The calibration parameters are not used in the 2D breathing estimation pipeline in this report. They are computed to enable future 3D reconstruction and multi-view fusion.

ROI Definition and Variants

To reduce the influence of background motion and irrelevant scene structure, our point tracking is restricted to a region of interest (ROI) derived from the subject's body. We consider three ROI configurations: **chest**, **abdomen**, and **a segmentation-based point selection**.

For the chest and abdomen-based ROIs, we use OpenPifPaf, a 2D human pose estimator that provides shoulder keypoints (Fig. 4). The shoulder landmarks define an upper-body reference line, from which we construct a rectangular ROI by extending the region downward. Chest and abdomen ROIs are obtained by selecting different vertical offsets based on this shoulder defined baseline. However, sometimes the model could fail at detecting shoulder keypoints. In that case, we fall into a hard-coded ROI computed in advance.

For the segmentation-based ROI, we compute a DeepLab person mask on the first frame and restrict the tracked point



(a) Olympus→Phone (b) Olympus→FLIR (c) Olympus→Gray

Fig. 3. Cross-view reprojection validation on a synchronized calibration frame. Filled markers: detected checkerboard corners in the target view. Cross markers: corners reprojected into the target view using estimated intrinsics/extrinsics and a pose estimated from a source view (olympus camera, cam0). Tight overlap indicates geometric consistency.

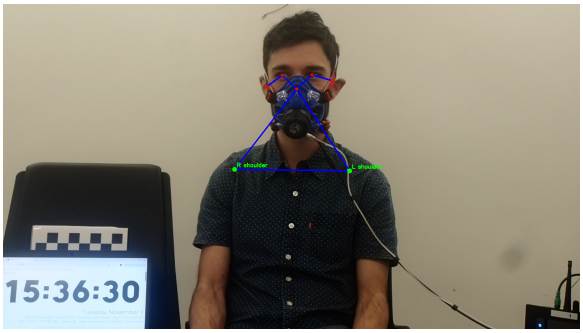


Fig. 4. OpenPifPaf keypoints detection.

set to points initialized inside the mask. This filtering suppresses contributions from background pixels and static scene elements, yielding motion signals that are more specific to the subject.

Appearance Normalization (CLAHE)

Respiratory motion produces small intensity changes that can be difficult to track on low-texture clothing. To improve the visibility of weak shading and fold patterns, we apply contrast-limited adaptive histogram equalization (CLAHE) [8] to the luminance channel in CIELAB space before point tracking. CLAHE increases local contrast while limiting noise amplification, making subtle local variations more visible.

CoTracker for Dense Point Tracking

Let T denote the number of frames in the sequence, $t \in \{0, \ldots, T-1\}$ the frame index, and N the number of tracked points initialized in the ROI.

Within each ROI, we initialize a regular grid of points and track them over time using CoTracker [1]. For each frame t and point i, the model outputs a 2D location $(x_{t,i}, y_{t,i})$ together with a visibility score $v_{t,i} \in [0, 1]$. We define a binary visibility mask $m_{t,i} = \mathbb{1}[v_{t,i} > 0.5]$ and ignore point measurements when $m_{t,i} = 0$.

Waveform Construction from Trajectories

We convert the dense trajectories into a 1D respiratory waveform by measuring displacement relative to the first frame:

$$\Delta x_{t,i} = x_{t,i} - x_{0,i}, \qquad \Delta y_{t,i} = y_{t,i} - y_{0,i}.$$

We evaluate two aggregation strategies:

- **Median vertical displacement:** for each frame t, we compute
$$s(t) = \text{median}_{i:\, m_{t,i}=1}\, \Delta y_{t,i}.$$
This estimator is robust to outliers caused by drifting points or occasional leakage outside the target region.
- **Dominant motion via PCA:** we form a feature matrix by concatenating the visible horizontal and vertical displacements, $\mathbf{X} \in \mathbb{R}^{T \times 2N}$, center each column, and extract the first principal component using SVD. The resulting score trajectory (PC1) is used as the waveform.

Signal Conditioning in the Breathing Band

The raw displacement waveform $s(t)$ may contain slow drift (e.g., posture changes) and high-frequency tracking noise. We therefore apply:

- detrending to remove low-frequency baseline variations,
- a low-order Butterworth band-pass filter (order 2) in a physiological breathing band (0.07–1.0 Hz),
- mild Gaussian smoothing to suppress frame-to-frame fluctuations while preserving respiratory periodicity

We additionally discard a short burn-in interval (2 s) to reduce boundary effects.

These described processing steps can be visualized in an example in Fig. 5.

Respiratory Rate Estimation

To obtain a robust RR estimate per phase, we compute Welch power spectra on multiple overlapping windows of the conditioned waveform. For each window, we select the dominant spectral peak within the breathing band and convert

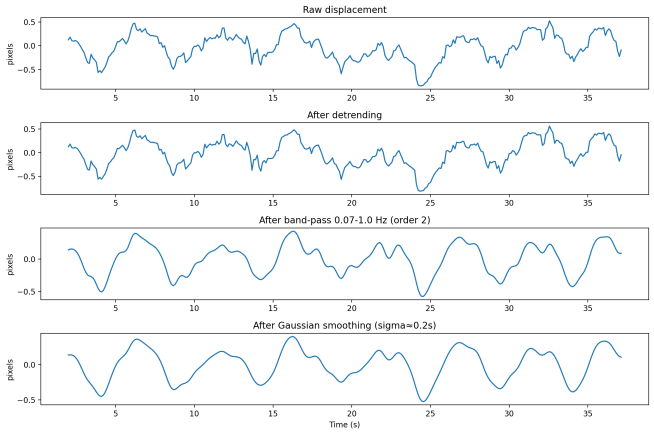


Fig. 5. Example waveform construction and conditioning for one phase segment.

it to breaths per minute. The final estimate aggregates window-level estimates by the median:

$$\widehat{Rf} = \text{median}_w \; 60 f_w^\star.$$

This windowed aggregation reduces sensitivity to local artifacts and transient motion unrelated to breathing.

Breath-Amplitude Proxy and VT Calibration

To obtain a breath-by-breath amplitude proxy, we detect inhalation peaks and exhalation valleys in the filtered waveform using prominence-based peak picking, and compute a peak-to-valley amplitude:

$$a_k = s(t_{\text{peak},k}) - s(t_{\text{valley},k}), \quad t_k \approx \tfrac{1}{2}(t_{\text{peak},k} + t_{\text{valley},k}) \, .$$

In other words, for each peak, we pair it with the most recent valley before, giving us one inhalation amplitude estimate per detected peak. Video breath event times are aligned to COSMED time via a small lag correction estimated from waveform cross-correlation. We match each reference timestamp to the nearest video event within a tolerance window to form paired samples (a_k, VT_k). A linear calibration model is then fit:

$$\widehat{VT} = \alpha a + \beta.$$

To mitigate optimistic bias due to temporal correlation, we use **blocked K-fold cross-validation** (contiguous folds) and report MAE, RMSE, and R^2 on held-out blocks.

Waveform Agreement with VT and VE

Beyond RR accuracy, we assess agreement between the extracted waveform and COSMED reference signals. After z-scoring, we compute:

- Pearson correlation between $s(t)$ and $VT_{\text{ref}}(t)$ / $VE_{\text{ref}}(t)$,
- an estimated best lag from cross-correlation (bounded), followed by resampling after applying the lag correction,
- magnitude-squared coherence in the breathing band, summarized by mean and peak coherence.

Coherence is computed using multiple overlapping segments to avoid the degenerate single-segment regime.

Fig. 6. Demo video of the tracking visualization, with dominent displacement.

V. Results

Experimental coverage

Our evaluation comprises **192 runs** in total, corresponding to

$$4 \text{ cameras} \times 2 \text{ subjects} \times 4 \text{ phases}$$

$$\times 3 \text{ ROIs} \times 2 \text{ aggregation methods} = 192.$$

We summarize performance using the **median** and **interquartile range (IQR)** to remain robust to occasional tracking failures and heavy-tailed error distributions.

ROI Visualization

Figures 7–9 illustrate the ROI definition and the resulting motion cues extracted across the four camera modalities.

Figure 7 shows the chest/abdomen ROI (red and blue boxes) for each viewpoint. Despite differences in resolution, field-of-view, and photometric appearance, the ROI is consistently centered on the torso region to capture respiration-induced motion while excluding most background content.

Figure 8 presents the segmentation-based ROI, where the estimated person mask is overlaid in red. This alternative ROI definition restricts the analysis to pixels belonging to the subject. The grayscale camera has no color cues and as DeepLabv3 pretraining expects natural RGB photos, it tends to produces many small mistakes in the person detection.

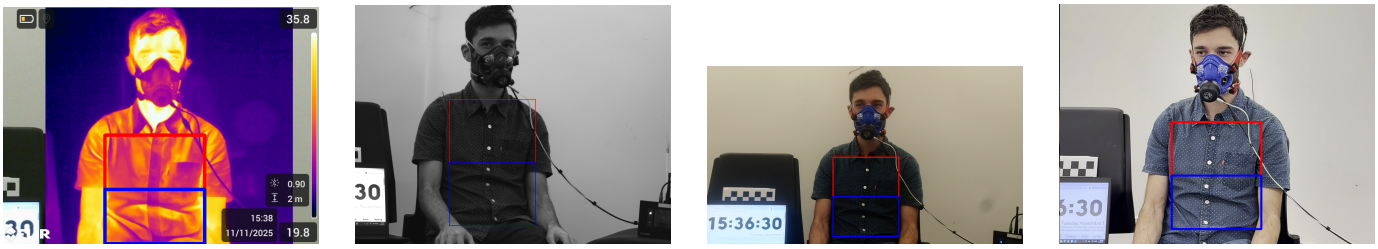
Figure 9 visualizes dense CoTracker point trajectories within the chest ROI over consecutive frames. The dense grid and the dominant displacement direction (yellow arrow) highlight coherent torso motion, consistent with the subtle, quasi-periodic movement associated with breathing.

Evaluation metrics

We report two complementary aspects of performance:
1. RR accuracy: absolute error $|e_{Rf}|$ between the video-derived respiratory rate $\widehat{Rf}$ and the mean COSMED respiratory rate Rf over the analyzed phase interval.
2. Waveform fidelity: magnitude-squared coherence between the video waveform $s(t)$ and COSMED reference waveforms $VT_{\text{ref}}(t)$ and $VE_{\text{ref}}(t)$, summarized within the breathing band (0.07–1.0 Hz).

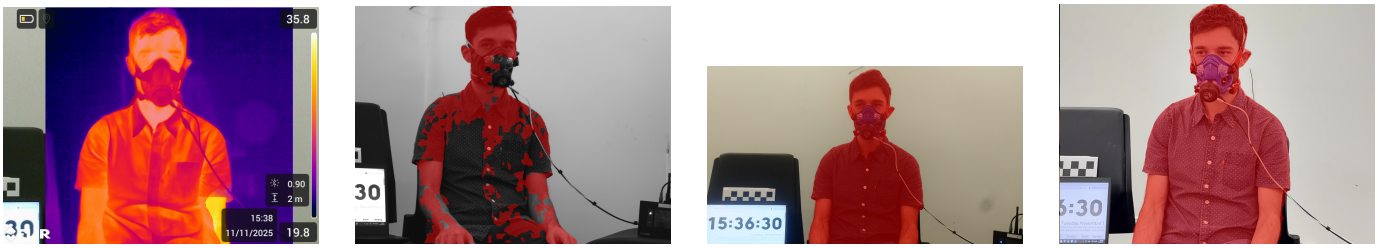
Aggregate performance across camera-specific runs

Table II summarizes performance across the full set of camera-specific runs (medians with interquartile ranges). Values are the median (IQR) across runs. Two robust trends emerge.



(a) FLIR (b) Gray (c) Olympus (d) iPhone

Fig. 7. Chest/abdomen ROI overlays across the four cameras.



(a) FLIR (b) Gray (c) Olympus (d) iPhone

Fig. 8. Segmentation-based ROI overlays across the four cameras.

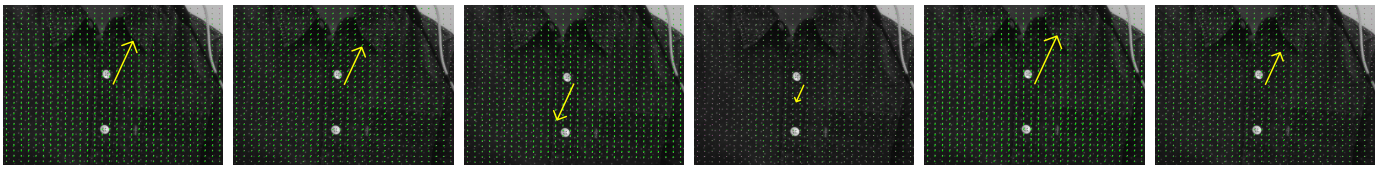


Fig. 9. Dense CoTracker point tracking within the chest ROI. The visualization highlights coherent torso motion across consecutive frames.

First, **chest ROIs yield the highest overall waveform agreement** with both VT and VE, consistent with chest motion being the most consistently respiration-locked source of displacement across viewpoints and modalities.

Second, **PCA improves RR slightly in the best case**, while the robust median often yields equal or better waveform coherence, reflecting a trade-off between frequency-based estimation (RR) and time/frequency consistency with reference waveforms.

TABLE II
PERFORMANCE BY ROI AND AGGREGATION METHOD ACROSS ALL RUNS.

ROI	Agg.	N	$\lvert e_{Rf} \rvert$ (bpm)	Coh_{VT}	Coh_{VE}
chest	median	32	1.99 (1.96)	0.34 (0.22)	0.37 (0.25)
chest	PCA	32	1.87 (1.76)	0.31 (0.18)	0.31 (0.17)
abdomen	median	32	2.15 (2.01)	0.33 (0.21)	0.34 (0.24)
abdomen	PCA	32	1.98 (1.92)	0.35 (0.23)	0.35 (0.20)
segmented	median	32	2.60 (2.29)	0.21 (0.17)	0.22 (0.16)
segmented	PCA	32	3.69 (4.89)	0.20 (0.16)	0.21 (0.13)

Cross-camera robustness analysis (median across cameras)

Because each subject has the same four phases recorded by all four cameras, we additionally evaluate a fused setting: for each configuration, we take the **median metric across the four cameras**.

Table III shows that **camera-median fusion improves waveform fidelity** (higher coherence medians) while leaving RR error broadly similar.

TABLE III
SAME SUMMARY AFTER FUSING CAMERAS: FOR EACH (SUBJECT, PHASE, ROI, METHOD), WE TAKE THE MEDIAN METRIC ACROSS THE FOUR CAMERAS AND THEN AGGREGATE ACROSS CONFIGURATIONS.

ROI	Agg.	N	$\lvert e_{Rf} \rvert$ (bpm)	Coh_{VT}	Coh_{VE}
chest	median	8	1.79 (1.48)	0.38 (0.13)	0.37 (0.13)
chest	PCA	8	1.71 (1.33)	0.32 (0.11)	0.32 (0.09)
abdomen	median	8	1.96 (1.54)	0.37 (0.15)	0.36 (0.16)
abdomen	PCA	8	1.90 (1.51)	0.36 (0.17)	0.35 (0.15)
segmented	median	8	2.35 (1.65)	0.23 (0.11)	0.23 (0.09)
segmented	PCA	8	3.22 (2.77)	0.20 (0.10)	0.23 (0.06)

Values are median (IQR). Camera fusion increases coherence medians relative to Table II, consistent with averaging out modality/viewpoint-specific tracking noise.

Recommended default settings

The final snapshot supports two practical defaults depending on the target:

TABLE IV
PER-CAMERA BEST CONFIGURATIONS FOR RR (MIN MEDIAN $|e_{Rf}|$). MEDIANS ARE TAKEN ACROSS SUBJECTS AND PHASES FOR EACH CAMERA.

Camera	ROI	Agg.	$\lvert e_{Rf}\rvert$ (bpm)
flir	chest	median	1.52
gray	chest	PCA	1.94
olympus	chest	PCA	1.73
phone	abdomen	PCA	2.62

TABLE V
PER-CAMERA BEST CONFIGURATIONS FOR WAVEFORM FIDELITY (MAX MEDIAN COH_{VT}). MEDIANS ARE TAKEN ACROSS SUBJECTS AND PHASES FOR EACH CAMERA.

Camera	ROI	Agg.	Coh_{VT}
flir	chest	median	0.42
gray	chest	median	0.39
olympus	chest	median	0.43
phone	abdomen	PCA	0.32

- **Waveform-first (best fidelity to VT/VE): chest ROI + median aggregation**. This maximizes the average coherence across VT and VE after camera fusion (Table III).
- **RR-first (best rate accuracy): chest ROI + PCA aggregation**. This yields the lowest median $|e_{Rf}|$ after camera fusion (Table III).

Per-camera best configurations

Although chest-focused ROIs dominate overall, optimal settings can vary by modality/viewpoint. Table IV reports, for each camera, the configuration that minimizes RR error and the configuration that maximizes waveform coherence (medians across subjects and phases).

Breath-level VT calibration (status)

Breath-level calibration metrics (blocked CV for $VT \approx a \cdot \text{proxy} + b$) are currently available only for a subset of runs that pass the inclusion criterion (at least 20 matched breaths). Across the snapshot, **24/192 runs** meet this criterion and the calibration is not yet reliably explaining VT variance across matched breaths in those runs. We therefore treat VT calibration as preliminary and focus the main "best setting" conclusions on RR and waveform fidelity.

VI. DISCUSSION

Why chest-focused ROIs are consistently strong

Across both camera-specific and camera-fused evaluations (Tables II–III), chest ROIs provide the most reliable waveform agreement with COSMED and competitive RR accuracy. This is consistent with the chest region concentrating respiration displacement while minimizing confusions from the abdomen boundary, clothing folds that move non-periodically, and partial inclusion of static background pixels near the torso contour. Because the cameras were fixed, the advantage of the chest ROI reflects improved signal-to-noise in the tracked point set.

Median vs. PCA aggregation: fidelity–rate trade-off

The observed trade-off is expected. The **median** displacement is not necessarily the best estimator but as long as more than half of the visible points carry respiration motion, the median suppresses outliers caused by drift, occlusions, and background leakage. This tends to preserve breathing-band structure and yields higher coherence, which is critical for downstream tasks that rely on waveform shape (e.g., breath event detection and amplitude proxies).

In contrast, **PCA** extracts the direction of maximum variance in the displacement field. When respiration dominates variance within the ROI, PCA sharpens the spectral peak and slightly improves RR (notably for chest ROIs). However, PCA can be distracted by competing components that explain more variance than respiration, such as slow posture drift or low-frequency non-respiratory deformation, and could reduce waveform agreement even if the dominant breathing frequency remains identifiable.

Practically, this supports using **chest+median** as the default for waveform-quality objectives and **chest+PCA** when RR is the principal target.

Why segmentation filtering underperforms in the current pipeline

The segmented condition is consistently weaker, particularly for RR (Tables II–III). A key issue is that the current segmentation behaves as a *hard gate* on tracked points. If the mask is imperfect (common under thermal imagery, grayscale contrast changes, or unusual framing), useful torso points may be excluded while residual background points remain. Additionally, because filtering is applied based on initialization, any systematic mask error persists throughout the run. This failure mode is consistent with a reduction in effective point count and a noisier aggregate signal, which harms both coherence and the robustness of PSD peak selection.

Why multi-camera median fusion improves coherence

Camera-median fusion increases coherence medians relative to camera-specific aggregation (compare Tables II and III). This is consistent with coherence being sensitive to modality/viewpoint-specific tracking artifacts that differ across cameras (texture richness, resolution, compression, thermal contrast). Taking the median across cameras reduces the influence of a single camera whose tracked waveform is contaminated by resolutions, orientation or background leakage, yielding a waveform that is more consistently respiration-locked. Importantly, RR is less sensitive to such distortions (it only requires a stable dominant frequency), explaining why RR changes only modestly under fusion.

Limitations and next steps

The conclusions above are stable at the level of RR/coherence trade-offs, but several limitations remain. First, segmentation is not yet optimized and likely penalizes the segmented condition. Second, ROI accuracy depends on pose estimation quality and ROI refinement (margins, exclusion

of arms/head, camera-specific scaling) may further improve robustness. Third, VT calibration requires improved breath matching and more runs with sufficient matched breaths.

Based on the final snapshot, the most impactful next steps are:

1) **Adopt chest+median as the waveform-default** and use chest+PCA only when RR is the single objective.
2) **Replace hard segmentation with soft weighting** and add temporal smoothing of masks.
3) **Exploit multi-camera fusion explicitly** for waveform-based tasks (breath events, amplitude proxies), given the clear coherence gain.
4) **Strengthen VT calibration** by improving event detection/matching, and expanding the proxy feature set beyond a single amplitude measure.

VII. Conclusion

Using dense tracking within pose-derived torso ROIs and classical signal processing, we obtain low-median RR errors (best: chest+PCA) while maintaining the highest waveform fidelity with robust aggregation (best: chest+median). Multi-camera median fusion improves waveform coherence, indicating that complementary viewpoints can stabilize respiration-locked motion extraction even when individual modalities are noisy. VT calibration remains preliminary and motivates further work on segmentation robustness, breath event matching, and richer proxies for volume estimation.

VIII. Future Work

With multi-view calibration in place, a natural next step is to move from per-camera 2D breathing estimates to a fused 3D representation of respiratory motion. This would enable improved robustness through view fusion and physiologically more interpretable amplitude measures.

Given synchronized videos and known camera intrinsics/extrinsics, we plan to extend the current pipeline in the following direction:

- **Cross-view correspondence for tracked points.** For each phase interval, dense point trajectories extracted in each view can be constrained by epipolar geometry [9] to identify candidate correspondences across cameras.
- **Triangulation and 3D trajectory estimation.** Once correspondences are established, we can triangulate matched points to obtain 3D trajectories over time, optionally refining them by minimizing reprojection error across all cameras.
- **3D breathing waveform extraction.** The resulting 3D point cloud motion can be aggregated into a respiratory waveform using robust statistics (e.g., median displacement magnitude) or a dominant 3D motion direction (3D PCA), analogous to our current 2D aggregation but in Euclidean space.

A 3D representation can provide amplitude features that are less view-dependent than 2D pixel motion, such as 3D displacement magnitude or expansion along an estimated chest-wall normal direction. We expect this to stabilize breath-level peak/valley detection and improve generalization of the linear calibration to VT and VE across viewpoints and modalities.

Code Availability

The code to reproduce the experiments and figures is available at this link (click).

AI Disclosure

This report was prepared with limited assistance from a generative AI tool (ChatGPT) for language editing, clarification of concepts, and for converting numerical results produced by the author's code into LaTeX table formatting. The AI tool was not used to generate or alter experimental results. All AI outputs were reviewed and verified against the original outputs by the author, who takes full responsibility for the final text, code, analysis, and results.

Appendix

Section IV defines the ROI variants used for tracking. Here (7, 8), we provide qualitative ROI overlays for each camera modality/viewpoint to illustrate typical framing differences and ROI localization behavior.

The pipeline follows Sec. IV. Here we document implementation-specific engineering choices (model caching, memory limits, run bookkeeping) and summarize hyperparameters (Table VI).

Software stack and compute backends

The pipeline is implemented in Python and relies on: PyTorch (model inference), OpenCV (image I/O + preprocessing), CoTracker3 (dense point tracking), OpenPifPaf (keypoint detection for ROI), torchvision DeepLabV3-ResNet50 (person segmentation), and SciPy (filtering, PSD, coherence, peak detection).

To avoid repeated initialization overhead, three heavy models are lazily instantiated and cached globally: OpenPifPaf predictor, DeepLabV3 model and CoTracker3 offline model.

Memory-safe CoTracker inference

CoTracker memory grows quickly with spatial resolution. To prevent out-of-memory failures, the ROI is resized so that its largest dimension does not exceed:

$$\texttt{MAX\_INFERENCE\_DIM} = 480 \text{ pixels.}$$

All ROI frames are pre-allocated into a single array `video_roi[T,H,W,3]` to avoid Python list append spikes, then converted to a tensor of shape:

$$(1, T, 3, H, W), \quad \text{normalized to } [0, 1].$$

We run CoTracker offline with `grid_size=30`.

Batch execution and outputs

Each run is identified by `run_id = camera_subject_take_ROI_PCA`. The batch loops over:

$$\texttt{ROI} \in \{\texttt{chest}, \texttt{abdomen}, \texttt{segmented}\},$$

$$\texttt{PCA} \in \{\texttt{False}, \texttt{True}\}.$$

A run is skipped if a done flag exists, a `results_ts.csv` exists, or the `run_id` already appears in the live report CSV. Each run writes:

- `breathing_waveform.csv` (raw + filtered waveform),
- `results_ts.csv` (COSMED + aligned video waveform),
- `breath_level.csv` (breath timestamps + VT + proxy + matching residuals),
- PNG figures (ROI preview, overlays, coherence, calibration plots),
- a global live report CSV aggregating metrics over all runs.

TABLE VI
KEY HYPERPARAMETERS USED IN THE BREATHING PIPELINE.

Component	Setting
ROI max inference size	`MAX_INFERENCE_DIM`=480 px
CoTracker sampling	`grid_size`=30
Breathing band	$[0.07, 1.0]$ Hz
Band-pass filter	Butterworth, order 2
Smoothing	Gaussian, $\sigma = 0.1 \cdot fps$
Burn-in removal	2 s
RR estimation	Welch PSD, win 20s, hop 2s (median)
Breath matching tolerance	± 2 s
Calibration evaluation	Blocked K-fold CV

This appendix describes the calibration script used to estimate intrinsics for each camera and extrinsics relative to a reference camera (cam0). The script assumes four folders containing synchronized selected frames of a checkerboard.

Frame synchronization

Frames are synchronized by matching basenames (filename without extension) across the four camera folders, retaining only common keys. This guarantees that each calibration index corresponds to the same time instant for all cameras.

Corner detection and intrinsics

For each synchronized frame, we detect a (10×7) internal-corner chessboard with subpixel refinement. Only frames where *all four* cameras successfully detect corners are kept.

Intrinsics are estimated independently per camera using `cv2.calibrateCamera`. We report: (i) OpenCV RMS reprojection error, and (ii) mean per-frame reprojection error (in pixels).

Checkerboard symmetry handling

A rectangular checkerboard has four valid corner-order symmetries: identity, horizontal flip, vertical flip, and 180° rotation. Since different cameras may observe the board with reversed ordering, we search these symmetries for each camera (except cam0) to enforce pose consistency.

For each frame and each symmetry, we estimate the checkerboard pose via `solvePnP`. We then compute the relative transform between cam0 and cam$c$ induced by the board poses. The best symmetry minimizes a cost that combines:

- angular deviation (degrees) from a running reference rotation,
- translation deviation (mm) from a running reference translation (scaled).

Pose-consistency outlier rejection

After symmetry selection, we apply outlier rejection by comparing each frame's relative pose against a robust per-camera median transform. Frames are removed if:

$$\Delta\theta > 6^\circ \quad \text{or} \quad \|\Delta t\| > 200 \text{ mm}.$$

This step reduces the effect of corner detection failures, motion blur, or partial-board views.

Extrinsics via stereo calibration with fixed intrinsics

For each camera cam$c$ (c=1,2,3), we estimate extrinsics relative to cam0 using:

`cv2.stereoCalibrate` with `CALIB_FIX_INTRINSIC`.

This optimizes the rigid transform $R_{0\to c}, t_{0\to c}$ while keeping intrinsics fixed. We save all parameters to a JSON file including:

- intrinsics: K, distortion coefficients, and image size per camera,
- extrinsics: $R_{0\to c}$, $t_{0\to c}$ (mm), and a 4x4 homogeneous transform matrix.

TABLE VII
KEY CALIBRATION SETTINGS.

Setting	Value
Checkerboard corners	$(10, 7)$ internal corners
Square size	24 mm
Corner refinement	`cornerSubPix`, window $(11, 11)$
Symmetry candidates	id, flip_h, flip_v, rot180
Outlier thresholds	6° and 200 mm
Extrinsics estimation	stereoCalibrate, FIX_INTRINSIC
Output	JSON (intrinsics + extrinsics)

REFERENCES

[1] N. Karaev, I. Rocco, B. Graham, N. Neverova, A. Vedaldi, and C. Rupprecht, "Cotracker: It is better to track together," in *Computer Vision – ECCV 2024*. Springer, 2024.

[2] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. T. Freeman, "Eulerian video magnification for revealing subtle changes in the world," *ACM Transactions on Graphics (TOG)*, vol. 31, no. 4, 2012.

[3] N. Wadhwa, M. Rubinstein, F. Durand, and W. T. Freeman, "Phase-based video motion processing," *ACM Transactions on Graphics*, vol. 32, no. 4, 2013.

[4] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI)*, 1981.
[5] Z. Teed and J. Deng, "Raft: Recurrent all-pairs field transforms for optical flow," in *Computer Vision – ECCV 2020*. Springer, 2020.
[6] C. Massaroni, D. Simões Lopes, D. Lo Presti, E. Schena, and S. Silvestri, "Contactless monitoring of breathing patterns and respiratory rate at the pit of the neck: A single camera approach," *Journal of Sensors*, vol. 2018, 2018.
[7] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000.
[8] K. Zuiderveld, "Contrast limited adaptive histogram equalization," in *Graphics Gems IV*. Academic Press, 1994.
[9] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*, 2nd ed. Cambridge University Press, 2004.