ViTaL-Diff: Video-Token Latent Diffusion for Contactless Respiratory Monitoring

Anonymous Authors[1]

Abstract

Contactless respiratory monitoring from video is a challenging inverse problem because breathing is observed only indirectly through subtle body and clothing motion, which can be corrupted by illumination, occlusion, pose variation, and non-respiratory movement. We propose **ViTaL-Diff**, a video-token latent diffusion framework that reconstructs respiratory waveforms from short upper-body videos without contact sensors at inference. ViTaL-Diff learns a compact respiratory latent space from belt-derived waveforms and trains a video-conditioned diffusion transformer to generate respiratory latent tokens from spatiotemporal video evidence. By modeling a distribution over plausible respiratory waveforms, the method supports both respiratory-rate estimation and uncertainty quantification. Across three datasets, including an in-house RGB dataset, AIR-125, and a low-light Sleep dataset, ViTaL-Diff achieves the lowest error among classical, deterministic, and diffusion baselines, with up to 28.9% mean absolute error (MAE) reduction over deterministic deep baselines and uncertainty estimates that identify visually ambiguous clips.

1. Introduction

Respiratory rate (RR) is a core vital sign for cardiopulmonary assessment, sleep monitoring, neonatal care, and remote health monitoring (Cretikos et al., 2008; Majumder et al., 2017; Massaroni et al., 2020). Although contact sensors and wearables support continuous RR measurement, they can be uncomfortable, motion-sensitive, and difficult to deploy long term (Massaroni et al., 2019a). Video-based monitoring offers a scalable contactless alternative by capturing respiration-induced cues from the chest, abdomen, shoulders, face, or clothing (Massaroni et al., 2019b; Queiroz et al., 2020).

Existing video-based RR methods use handcrafted motion extraction or learned spatiotemporal representations. Classical approaches rely on optical flow, frame differencing, phase-based analysis, motion magnification, and filtered ROI signals (Siam et al., 2020; Tan et al., 2010; Massaroni et al., 2019a; Manne et al., 2023), but are fragile under illumination changes, occlusion, weak respiratory motion, and non-respiratory movement. Recent deep models learn video time-series features from RGB, NIR, thermal, or flow-based representations (Khanam et al., 2021; Sakib et al., 2025c; Figueroa et al., 2019; Hasan et al., 2025), while remote-physiology models such as DeepPhys (Chen & McDuff, 2018), TS-CAN (Liu et al., 2020), PhysNet (Yu et al., 2019), and EfficientPhys (Liu et al., 2023) show that spatiotemporal networks can recover physiological signals from video. However, most methods remain deterministic, mapping each clip to a single waveform or RR estimate even when visual evidence is weak or ambiguous. Contactless RR estimation is therefore an ill-posed *inverse problem*: different respiratory waveforms can produce similar visual motion, and nuisance motion can mimic breathing. Diffusion models are well suited to this setting because they learn conditional time-series distributions rather than point estimates (Ho et al., 2020). They have shown strong results in speech generation (Ghosal et al., 2023), time-series generation (Yuan & Qiao, 2024), imputation (Tashiro et al., 2021), biomedical denoising (Liu et al., 2025), and respiratory waveform estimation from PPG (Miao et al., 2025). Yet prior respiratory diffusion methods mainly use contact-sensor inputs or operate directly in waveform space.

We propose ViTaL-Diff, a video-token latent diffusion framework for respiratory waveform reconstruction from upper-body video time series. It learns a compact respiratory latent space from belt-derived waveforms and generates respiratory latent tokens from spatiotemporal video evidence, enabling RR estimation and uncertainty quantification.

Our contributions are summarized as follows:

- We formulate contactless respiratory monitoring from video time series as conditional latent generative mod-

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

eling for uncertainty-aware reconstruction.

- We introduce ViTaL-Diff, a video-conditioned latent diffusion transformer for generating respiratory time-series tokens from visual evidence.
- We evaluate on an in-house RGB dataset, AIR-125, and a low-light Sleep dataset, showing consistent gains over classical, deterministic, and waveform-space diffusion baselines.

2. Problem Setup

We study contactless respiratory waveform reconstruction from short upper-body video clips. During training, each video clip $V = \{F_1, \dots, F_N\}$ is paired with a synchronized respiratory-belt waveform $y \in \mathbb{R}^T$; at inference time, only V is available. Given paired samples $\mathcal{D} = \{(V_i, y_i)\}_{i=1}^{N_d}$, the goal is to learn a conditional generative model $p_\theta(y \mid V)$ that reconstructs a physiologically plausible respiratory waveform from video alone. This mapping is ambiguous under weak respiratory motion and visual artifacts. We therefore adopt a generative formulation rather than a single deterministic estimate. ViTaL-Diff learns a latent conditional model $p_\theta(z_0 \mid V)$, where the belt waveform is encoded as respiratory latent tokens $z_0 = E_R(y) \in \mathbb{R}^{L\times d}$, and the waveform is recovered by a decoder D_R. This latent formulation reduces diffusion dimensionality and enables uncertainty estimation through sample variability.

3. Methodology

ViTaL-Diff first learns a compact latent representation of respiratory waveforms, then trains a video-conditioned diffusion model to generate respiratory latent time-series tokens from visual evidence. As shown in Figure 1, the framework contains three modules: (i) a respiratory waveform tokenizer, (ii) a video evidence encoder, and (iii) a video-conditioned latent diffusion transformer.

3.1. Respiratory Waveform Tokenizer

The tokenizer maps each belt-derived respiratory waveform into a compact sequence of L latent tokens, each with dimension d:

$$z_0 = E_R(y), \qquad z_0 \in \mathbb{R}^{L\times d}. \tag{1}$$

The decoder reconstructs the waveform from this latent sequence, $\hat{y}_{tok} = D_R(z_0)$. The tokenizer is trained with

$$\mathcal{L}_{tok} = \|y - D_R(E_R(y))\|_1 . \tag{2}$$

After training, E_R and D_R are frozen. The encoder provides clean latent targets for diffusion training, while the decoder maps generated latent tokens back to waveform space during inference.

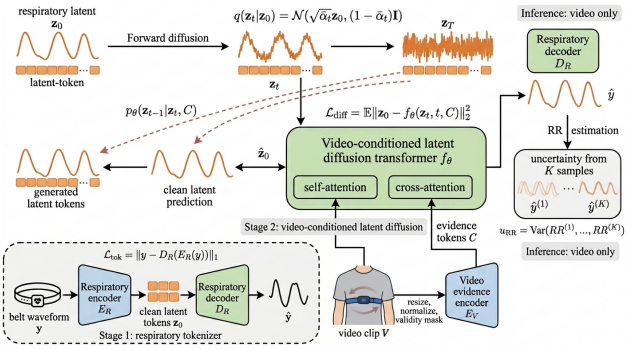


Figure 1. Overview of **ViTaL-Diff**: video evidence tokens condition latent diffusion over respiratory waveform tokens.

3.2. Video Evidence Encoder

The video clip is treated as a visual time series and divided into spatiotemporal tubelets, which are embedded as visual tokens and processed by a video encoder:

$$C = E_V(V), \qquad C \in \mathbb{R}^{M\times d}. \tag{3}$$

Here, C is a sequence of M time-ordered video evidence tokens projected to the same token dimension as the respiratory latent space. Temporal positional embeddings preserve frame order, and validity masks suppress missing or corrupted frames. The evidence tokens C are the only conditioning input to the diffusion model, allowing ViTaL-Diff to learn respiratory cues directly from video time series rather than frequency features.

3.3. Video-Conditioned Latent Diffusion

The diffusion model is trained in the respiratory latent space. For each pair (V, y), the frozen respiratory encoder produces $z_0 = E_R(y)$, and the video encoder produces $C = E_V(V)$. The forward process corrupts z_0 with Gaussian noise:

$$q(z_t \mid z_0) = \mathcal{N}\left(z_t; \sqrt{\bar{\alpha}_t}z_0, (1-\bar{\alpha}_t)I\right), \tag{4}$$

where $\bar{\alpha}_t = \prod_{s=1}^{t}(1-\beta_s)$. Equivalently,

$$z_t = \sqrt{\bar{\alpha}_t}z_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I). \tag{5}$$

The denoising network predicts the clean latent tokens from noisy tokens, timestep, and video evidence:

$$\hat{z}_{0,\theta} = f_\theta(z_t, t, C). \tag{6}$$

We use clean latent prediction instead of noise-only prediction to better preserve temporal structure. The diffusion objective is

$$\mathcal{L}_{diff} = \mathbb{E}\left[\|z_0 - f_\theta(z_t, t, C)\|_2^2\right]. \tag{7}$$

Training is therefore two-stage: first train the tokenizer with $\mathcal{L}_{tok}$, then freeze it and train the video-conditioned diffusion model with $\mathcal{L}_{diff}$.

3.4. Latent Diffusion Transformer

The denoising network f_θ is a conditional transformer over respiratory latent time-series tokens. Each block first applies self-attention over the noisy latent sequence,

$$H' = \mathrm{SelfAttn}(H), \tag{8}$$

followed by cross-attention from respiratory latent tokens to video evidence tokens,

$$H'' = \mathrm{CrossAttn}(H', C). \tag{9}$$

Self-attention captures temporal dependencies across the respiratory time series, while cross-attention aligns each respiratory latent token with relevant time-ordered video evidence.

3.5. Robust Conditioning and Uncertainty Estimation

To improve robustness to weak motion, occlusion, illumination variation, and dropped frames, we apply evidence dropout during training:

$$\tilde{C} = \mathrm{DropEvidence}(C, \rho), \tag{10}$$

where ρ controls the fraction of evidence tokens removed. With probability p_{uncond}, the evidence tokens are replaced by a learned null condition, enabling classifier-free guidance:

$$\hat{z}_0 = (1 + w) f_\theta(z_t, t, C) - w f_\theta(z_t, t, \varnothing). \tag{11}$$

The scalar w controls the strength of video conditioning.

At inference time, only video is used. We encode V into evidence tokens C, sample $z_T \sim \mathcal{N}(0, I)$, and iteratively denoise to obtain $\hat{z}_0$. The waveform is reconstructed as

$$\hat{y} = D_R(\hat{z}_0). \tag{12}$$

RR is computed from $\hat{y}$ using a fixed post-processing estimator such as peak counting or dominant-frequency estimation. Since the model is generative, we draw K samples and compute the posterior mean waveform,

$$\bar{y} = \frac{1}{K} \sum_{k=1}^{K} \hat{y}^{(k)}. \tag{13}$$

RR uncertainty is estimated as:

$$u_{RR} = \mathrm{Var}\left(RR^{(1)}, \ldots, RR^{(K)}\right). \tag{14}$$

A high value of u_{RR} indicates weak visual evidence; thus, ViTaL-Diff provides both a respiratory estimate and a reliability signal.

4. Experiments

4.1. Datasets and Preprocessing

We evaluate ViTaL-Diff on three datasets: an in-house 50-video RGB dataset, AIR-125 (Manne et al., 2023), and a Sleep dataset (Hu et al., 2018). The in-house dataset was collected under an institutionally approved IRB protocol and contains synchronized upper-body RGB recordings with a Vernier respiration belt (VER) used to obtain ground-truth respiratory waveforms. AIR-125 contains 125 pediatric videos from diverse real-world settings, with RR ranging from 18 to 42 bpm and varied resolutions and frame rates. The Sleep dataset contains 28 synchronized IR/NIR recordings from 12 adults under low-light sleep conditions, enabling evaluation in low-motion nighttime settings. Across datasets, videos are paired with available respiratory references. For the in-house dataset, the Vernier belt signal is temporally aligned to each video window and normalized per recording. Frames are decoded, invalid frames are removed, resized, normalized, and segmented into 6-second windows with 3-second overlap, producing paired samples (V_i, y_i), where y_i is the ground-truth waveform or RR reference. Classical ROI, optical-flow, FFT, and peak-based processing are used only for baselines and diagnostics. Additional dataset, Vernier belt, video–belt alignment, and diagnostic motion-signal processing details are provided in Sections A.1 to A.4.

4.2. Baselines and Metrics

We compare ViTaL-Diff with five baselines: (i) Peak/FFT estimates RR from filtered video-derived motion signals using peak counting or dominant-frequency estimation; (ii) E2RespUNet (Sakib et al., 2025b) and (iii) RespFormer (Sakib et al., 2025a) represent recent contactless respiratory monitoring models for waveform/RR estimation; (iv) latent regression uses the same respiratory tokenizer as ViTaL-Diff but predicts z_0 directly from video evidence without diffusion; and (v) waveform-space diffusion applies conditional diffusion directly to respiratory waveforms instead of latent respiratory tokens. We report RR estimation performance using MAE, RMSE, and Pearson correlation r across all datasets. When reference waveforms are available, we additionally report waveform MAE and waveform correlation. For uncertainty evaluation, ViTaL-Diff samples K waveforms per clip and computes u_{RR}, the variance of sampled RR estimates. We further assess reliability by reporting MAE after rejecting high-u_{RR} clips.

4.3. Implementation Details

ViTaL-Diff is trained in two stages. We train the respiratory tokenizer with $\mathcal{L}_{tok}$ for 50 epochs, freeze it, and then train the video-conditioned latent diffusion transformer with

Table 1. Respiratory-rate MAE across datasets. ViTaL-Diff consistently outperforms classical, deterministic, and diffusion baselines.

Method	Dataset		
	In-house RGB	AIR-125	Sleep
Peak/FFT	4.6	6.1	5.4
E2RespUNet	3.8	4.9	4.1
RespFormer	3.3	4.8	4.5
Latent regression	3.6	4.4	4.2
Waveform-space diffusion	3.2	4.7	4.1
ViTaL-Diff	**2.7**	**4.2**	**3.5**

$\mathcal{L}_{diff}$ for 100 epochs. We use AdamW with learning rate 1×10^{-4}, weight decay 1×10^{-4}, cosine decay, and batch size 8. Since diffusion is performed in a compact respiratory latent space, we use 200 training timesteps and 20 DDIM sampling steps. Evidence dropout uses $\rho = 0.15$, and classifier-free conditioning dropout uses $p_{\text{uncond}} = 0.1$. At inference, only video is used; latent tokens are decoded into waveforms, and RR is computed from the posterior mean using $K = 10$ samples per clip. Experiments are run on an NVIDIA H200 GPU server, with subject-independent splits when subject identifiers are available.

4.4. Results and Discussion

Table 1 reports respiratory-rate MAE across the three datasets. ViTaL-Diff achieves the lowest error in every setting, showing consistent gains over classical signal processing, deterministic neural models, and diffusion-based variants. Against RespFormer, a strong deterministic baseline, ViTaL-Diff reduces MAE by 18.2%, 12.5%, and 22.2% on the in-house RGB, AIR-125, and Sleep dataset, respectively. Compared with E2RespUNet, ViTaL-Diff reduces MAE by 28.9%, 14.3%, and 14.6%. These results indicate that modeling respiratory dynamics as a conditional distribution is more effective than producing a single deterministic estimate from video.

The controlled variants clarify the source of the improvement. Waveform-space diffusion benefits from probabilistic generation, but it denoises raw respiratory samples directly. ViTaL-Diff instead denoises compact respiratory tokens, reducing MAE by 15.6%, 10.6%, and 14.6% over waveform-space diffusion. This supports moving diffusion from raw waveform samples to a learned respiratory token space. Latent regression uses the same tokenizer but removes diffusion sampling, reducing the model to a deterministic video-to-latent mapping. ViTaL-Diff improves over this variant by 25.0%, 4.5%, and 16.7%, showing that tokenization alone is insufficient and that conditional sampling helps under ambiguous visual evidence. The largest relative improvements occur over the Peak/FFT baseline, where ViTaL-Diff reduces MAE by 41.3%, 31.1%, and 35.2%. This is expected:

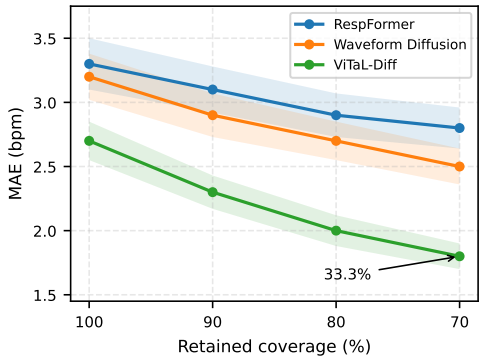


Figure 2. Uncertainty–coverage analysis. ViTaL-Diff is compared with deterministic and generative baselines; MAE decreases as high-u_{RR} clips are rejected.

frequency-domain estimates can be unstable when respiratory motion is weak, mixed with non-respiratory movement, or affected by harmonic ambiguity. In contrast, ViTaL-Diff learns visual respiratory evidence directly from video tokens and constrains generation through a respiratory latent prior. The consistent improvement across RGB, pediatric, and low-light sleep recordings suggests that the learned latent formulation is robust across different sensing conditions.

Figure 2 evaluates whether the generative model provides a useful reliability signal. We compare ViTaL-Diff with RespFormer and waveform-space diffusion, representing the strongest deterministic and closest generative baselines. As clips with high sampled RR variance u_{RR} are rejected, MAE decreases monotonically. On the in-house RGB dataset, rejecting the top 20% most uncertain clips lowers ViTaL-Diff MAE from 2.7 bpm to 2.0 bpm, a 25.9% reduction; rejecting the top 30% lowers MAE to 1.8 bpm, a 33.3% reduction. Thus, sample variability is not merely a byproduct of diffusion sampling, but a meaningful indicator of weak-motion, occluded, or visually ambiguous clips. Together, these results show that latent diffusion improves RR estimation, outperforms waveform-space diffusion, and yields a practical uncertainty signal for contactless monitoring.

5. Conclusion

We presented ViTaL-Diff, a latent diffusion framework for contactless respiratory waveform reconstruction from upper-body video time series. By generating respiratory latent tokens conditioned on video evidence, ViTaL-Diff improves RR estimation across RGB, pediatric, and sleep settings while avoiding handcrafted motion features. Its sample variability provides a practical reliability measure for ambiguous clips, which is important for deployment in uncontrolled environments. Future work will extend toward broader clinical validation, real-time deployment, and robustness across diverse cameras, lighting conditions, and patient populations.

References

Vernier go direct respiration belt. https://www.vernier.com/product/go-direct-respiration-belt/.

Chen, W. and McDuff, D. Deepphys: Video-based physiological measurement using convolutional attention networks. In *Proceedings of the european conference on computer vision (ECCV)*, pp. 349–365, 2018.

Cretikos, M. A., Bellomo, R., Hillman, K., Chen, J., Finfer, S., and Flabouris, A. Respiratory rate: the neglected vital sign. *Medical Journal of Australia*, 188(11):657–659, 2008.

Figueroa, I. R. A., Nuño, J. V. M., and Mendizabal-Ruiz, E. G. Remote optical estimation of respiratory rate based on a deep learning human pose detector. In *Latin American Conference on Biomedical Engineering*, pp. 234–241. Springer, 2019.

Ghosal, D., Majumder, N., Mehrish, A., and Poria, S. Text-to-audio generation using instruction guided latent diffusion model. In *Proceedings of the 31st ACM international conference on multimedia*, pp. 3590–3598, 2023.

Hasan, Z., Ahmed, M., Sakib, S., Chugh, S., Khan, M. A., Zaher MD Faridee, A., and Roy, N. Rrpips: Respiratory waveform reconstruction using persistent independent particles tracking from video. In *2025 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, pp. 13–24, 2025.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Hu, M., Zhai, G., Li, D., Fan, Y., Duan, H., Zhu, W., and Yang, X. Combination of near-infrared and thermal imaging techniques for the remote and simultaneous measurements of breathing and heart rates under sleep situation. *PloS one*, 13(1):e0190466, 2018.

Khanam, F.-T.-Z., Perera, A. G., Al-Naji, A., Gibson, K., and Chahl, J. Non-contact automatic vital signs monitoring of infants in a neonatal intensive care unit based on neural networks. *Journal of Imaging*, 7(8):122, 2021.

Langley, P. Crafting papers on machine learning. In Langley, P. (ed.), *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, pp. 1207–1216, Stanford, CA, 2000. Morgan Kaufmann.

Liu, X., Fromm, J., Patel, S., and McDuff, D. Multi-task temporal shift attention networks for on-device contactless vitals measurement. *Advances in Neural Information Processing Systems*, 33:19400–19411, 2020.

Liu, X., Hill, B., Jiang, Z., Patel, S., and McDuff, D. Efficientphys: Enabling simple, fast and accurate camera-based cardiac measurement. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 5008–5017, 2023.

Liu, Y.-T., Wang, K.-C., Chao, R., Siniscalchi, S. M., Yeh, P.-C., and Tsao, Y. Msemg: Surface electromyography denoising with a mamba-based efficient network. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Majumder, S., Mondal, T., and Deen, M. J. Wearable sensors for remote health monitoring. *Sensors*, 17(1):130, 2017.

Manne, S. K. R., Zhu, S., Ostadabbas, S., and Wan, M. Automatic infant respiration estimation from video: A deep flow-based algorithm and a novel public benchmark. In *International Workshop on Preterm, Perinatal and Paediatric Image Analysis*, pp. 111–120. Springer, 2023.

Massaroni, C., Nicolò, A., Lo Presti, D., Sacchetti, M., Silvestri, S., and Schena, E. Contact-based methods for measuring respiratory rate. *Sensors*, 19(4):908, 2019a.

Massaroni, C., Schena, E., Silvestri, S., and Maji, S. Comparison of two methods for estimating respiratory waveforms from videos without contact. In *2019 IEEE International Symposium on Medical Measurements and Applications (MeMeA)*, pp. 1–6. IEEE, 2019b.

Massaroni, C., Nicolo, A., Sacchetti, M., and Schena, E. Contactless methods for measuring respiratory rate: A review. *IEEE Sensors Journal*, 21(11):12821–12839, 2020.

Miao, Y., Chen, Z., Li, C., and Mandic, D. P. Respdiff: An end-to-end multi-scale rnn diffusion model for respiratory waveform estimation from ppg signals. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Queiroz, L., Oliveira, H., Yanushkevich, S., and Ferber, R. Video-based breathing rate monitoring in sleeping subjects. In *2020 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 2458–2464. IEEE, 2020.

Sakib, S., Shinde, G., Chugh, S., Anwar, M. S., and Roy, N. Respformer: A motion-guided temporal-frequency multimodal fusion transformer for contactless respiratory monitoring. In *2025 International Conference on Machine Learning and Applications (ICMLA)*, pp. 631–638. IEEE, 2025a.

Sakib, S., Shinde, G., Dev, E., and Roy, N. E2respunet: End-to-end respiratory signal reconstruction and rate prediction using a unified attention-enhanced u-net. In *2025 IEEE International Conference on Smart Computing (SMARTCOMP)*, pp. 154–161, 2025b. doi: 10.1109/SMARTCOMP65954.2025.00068.

Sakib, S., Shinde, G., Dev, E., and Roy, N. E2respunet: End-to-end respiratory signal reconstruction and rate prediction using a unified attention-enhanced u-net. In *2025 IEEE International Conference on Smart Computing (SMARTCOMP)*, pp. 154–161. IEEE, 2025c.

Siam, A. I., El-Bahnasawy, N. A., El Banby, G. M., Abou Elazm, A., and Abd El-Samie, F. E. Efficient video-based breathing pattern and respiration rate monitoring for remote health monitoring. *JOSA A*, 37(11):C118–C124, 2020.

Tan, K. S., Saatchi, R., Elphick, H., and Burke, D. Real-time vision based respiration monitoring system. *CSNDSP 2010*, pp. 770, 2010. doi: 10.1109/CSNDSP.2010.

Tashiro, Y., Song, J., Song, Y., and Ermon, S. Csdi: Conditional score-based diffusion models for probabilistic time series imputation. *Advances in neural information processing systems*, 34:24804–24816, 2021.

Yu, Z., Li, X., and Zhao, G. Remote photoplethysmograph signal measurement from facial videos using spatio-temporal networks. In *30th British Machine Visison Conference: BMVC 2019. 9th-12th September 2019, Cardiff, UK*. The British Machine Vision Conference (BMVC), 2019.

Yuan, X. and Qiao, Y. Diffusion-ts: Interpretable diffusion for general time series generation. *arXiv preprint arXiv:2403.01742*, 2024.

A. Appendix

A.1. In-House Dataset Construction

We developed an in-house video-based respiratory dataset to reflect both clinical and remote monitoring conditions. The dataset currently contains 50 RGB videos collected from 10 adult volunteers, with each recording lasting approximately 2.5–3.5 minutes at 1920×1080 resolution and 30 fps. Recordings include spontaneous breathing and paced breathing patterns, including slow, regular, and fast breathing. Videos were collected under multiple postures and viewpoints, including front-facing, back-facing, oblique, overhead, and supine views. The dataset also includes variation in lighting conditions, background motion, and environmental noise.

Participants rested for five minutes between recording sessions to reduce fatigue and carry-over effects. The cohort includes variation in age, gender, and skin tone. Data collection was conducted under an institutionally approved IRB protocol, and all participants provided informed consent with privacy and data-protection provisions. We are continuing to expand the dataset with additional subjects, postures, viewpoints, and environmental conditions. After de-identification and institutional approval, we plan to release the dataset to support reproducible research in contactless respiratory monitoring.

Ground-truth respiratory waveforms were collected using a Vernier Go Direct Respiration Belt, which measures breathing-induced pressure changes. The belt signal was synchronized with the video stream and used as the reference respiratory waveform. Reference respiratory rates were derived from the belt waveform using peak detection in 30-second sliding windows advanced every 10 seconds.

A.2. Belt-Signal Processing and Reference RR Construction

Let $s(t)$ denote the raw Vernier belt signal. The signal is first temporally aligned with video timestamps. To reduce subject-specific amplitude variation, the belt waveform is normalized within each recording:

$$\tilde{s}(t) = \frac{s(t) - \mu_s}{\sigma_s + \epsilon}, \tag{15}$$

where μ_s and σ_s are the mean and standard deviation of the belt signal, and ϵ prevents numerical instability.

To remove slow baseline drift and high-frequency noise, the normalized signal is filtered within the respiratory band:

$$s_f(t) = \mathrm{Bandpass}(\tilde{s}(t); f_{\min}, f_{\max}), \tag{16}$$

where $f_{\min}$ and $f_{\max}$ define the plausible breathing-frequency range. Peaks are detected from $s_f(t)$. For a window of length W, respiratory rate is computed as,

$$RR = \frac{60}{W}|\mathcal{P}_W|, \tag{17}$$

where $|\mathcal{P}_W|$ is the number of detected inhalation peaks in the window. In our processing, $W = 30$ seconds and the window advances every 10 seconds. Each 6-second video segment is assigned the nearest belt-derived RR value for segment-level evaluation. For waveform-supervised training, the aligned belt waveform segment is used directly:

$$y_i = s_f(t_i : t_i + T), \tag{18}$$

where T is the duration of the waveform segment aligned with video clip V_i.

A.3. Video Processing and Segment Construction

Each video is paired with its corresponding Vernier belt file using the recording identifier. Frames are decoded sequentially at 30 fps, invalid or unreadable frames are removed, and the remaining frames are resized and normalized. Each recording is segmented into 6-second windows with 3-second overlap, producing paired video–waveform examples:

$$(V_i, y_i), \tag{19}$$

where V_i is an upper-body video time series and y_i is the synchronized belt-derived respiratory waveform segment. This window length preserves local respiratory dynamics while keeping the video input short enough for efficient video-token modeling.

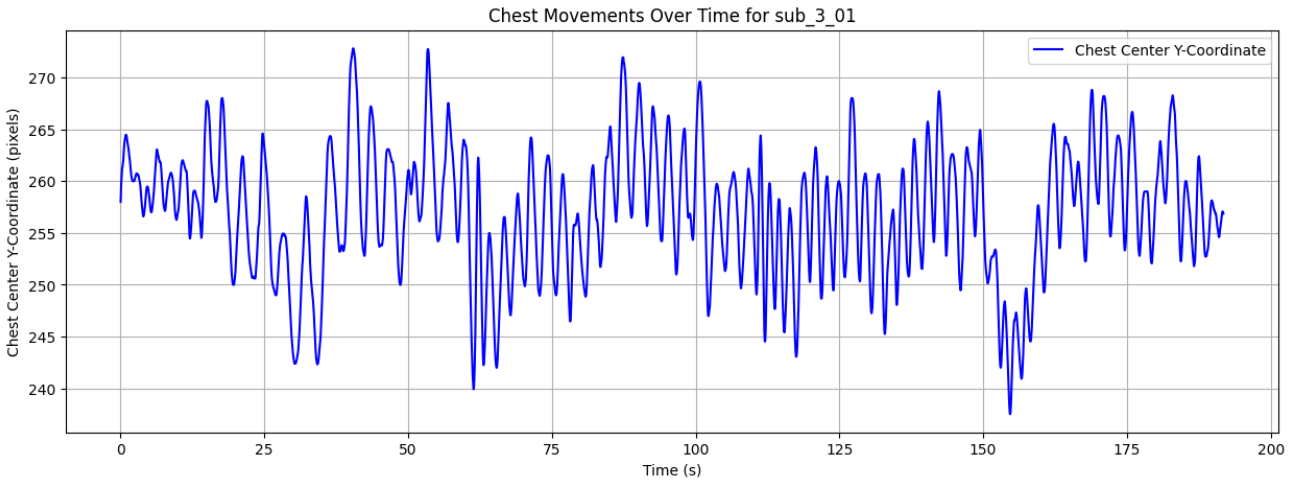


Figure 3. Example chest-motion trajectory from the in-house RGB dataset. The vertical chest-center coordinate varies over time due to breathing-induced upper-body motion and occasional non-respiratory movement.

A.4. Diagnostic Motion-Signal Construction

In addition to belt-based reference construction, we extract diagnostic video motion signals to inspect respiratory visibility and to implement classical baselines. For ROI-center tracking, the vertical chest displacement signal is represented as

$$m_y(t) = c_y(t), \tag{20}$$

where $c_y(t)$ is the vertical coordinate of the tracked chest-region center at time t. Since respiration induces cyclic chest and abdomen displacement, $m_y(t)$ often contains respiratory periodicity. Figure 3 shows an example chest-center trajectory over time, where periodic oscillations correspond to respiratory motion and abrupt deviations may indicate body movement or tracking noise. For frame-difference or optical-flow-based motion, the motion magnitude at frame t is computed over a

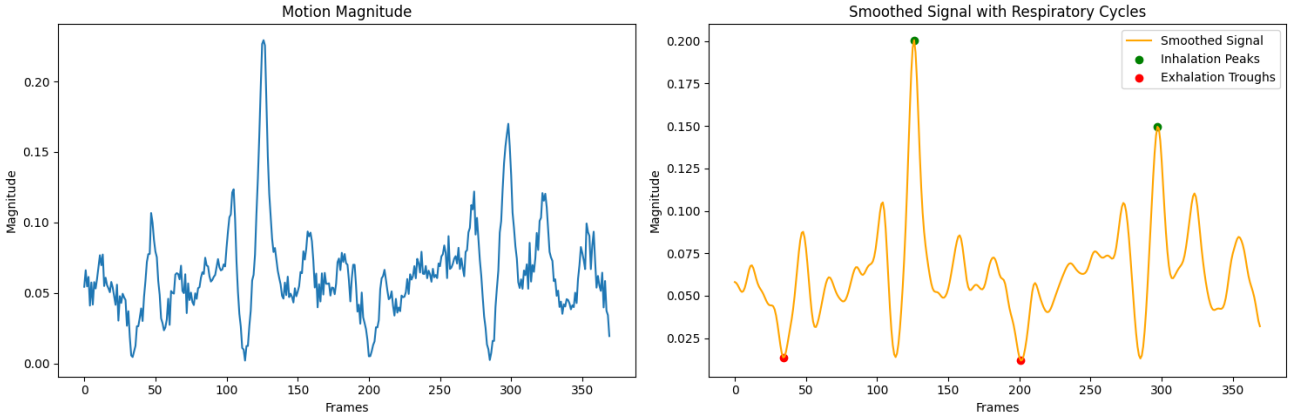


Figure 4. Diagnostic motion-signal processing for classical baselines. The raw motion magnitude is smoothed to reveal respiratory cycles, and peaks/troughs are detected to estimate respiratory rate.

chest ROI Ω:

$$m(t) = \frac{1}{|\Omega|} \sum_{p \in \Omega} \sqrt{u_t(p)^2 + v_t(p)^2}, \tag{21}$$

where $u_t(p)$ and $v_t(p)$ are the horizontal and vertical motion components at pixel p. The resulting one-dimensional motion signal is smoothed using a moving average or low-pass filter:

$$\bar{m}(t) = \frac{1}{K} \sum_{k=0}^{K-1} m(t-k). \tag{22}$$

Respiratory peaks and troughs are detected from $\bar{m}(t)$. Figure 4 illustrates this diagnostic signal-processing pipeline: the raw motion magnitude is smoothed to reveal respiratory cycles, and detected peaks/troughs are used for peak-based respiratory-rate estimation.

A.5. Frequency-Domain RR Estimation for Classical Baselines

For the Peak/FFT baseline, a one-dimensional motion signal $m(t)$ is filtered within the respiratory band. The dominant frequency is estimated from the Fourier magnitude spectrum:

$$f^\star = \arg \max_{f \in [f_{\min}, f_{\max}]} |\mathcal{F}\{m(t)\}(f)|, \tag{23}$$

where $\mathcal{F}\{\cdot\}$ denotes the Fourier transform. The respiratory rate is computed as:

$$RR_{\mathrm{FFT}} = 60 f^\star. \tag{24}$$

Peak-based RR is computed from the number of detected respiratory peaks:

$$RR_{\mathrm{peak}} = \frac{60}{W} N_{\mathrm{peaks}}, \tag{25}$$

where N_{peaks} is the number of peaks in a W-second window.

A.6. Algorithm

Algorithm 1 ViTaL-Diff: Training and Inference

Input: Paired data $\mathcal{D} = \{(V_i, y_i)\}_{i=1}^N$, respiratory tokenizer (E_R, D_R), video encoder E_V, denoising transformer f_θ.
Stage 1: Learn respiratory latent space
for each minibatch of waveforms y **do**
  Encode respiratory tokens: $z_0 = E_R(y)$.
  Reconstruct waveform: $\hat{y} = D_R(z_0)$.
  Update E_R, D_R using $\mathcal{L}_{tok} = \|y - D_R(E_R(y))\|_1$.
end for
Freeze E_R and D_R.
Stage 2: Train video-conditioned latent diffusion
for each minibatch (V, y) **do**
  Compute clean respiratory tokens: $z_0 = E_R(y)$.
  Compute video evidence tokens: $C = E_V(V)$.
  Sample t and $\epsilon \sim \mathcal{N}(0, I)$.
  Corrupt respiratory tokens:

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon.$$

  Apply evidence dropout to C, and optionally replace C with $\varnothing$ for classifier-free guidance.
  Predict clean tokens: $\hat{z}_{0,\theta} = f_\theta(z_t, t, C)$.
  Update E_V, f_θ using

$$\mathcal{L}_{diff} = \mathbb{E}\left[\|z_0 - f_\theta(z_t, t, C)\|_2^2\right].$$

end for
Inference: video only
Encode video evidence $C = E_V(V)$.
for $k = 1, \ldots, K$ **do**
  Sample $z_T^{(k)} \sim \mathcal{N}(0, I)$ and denoise with DDIM to obtain $\hat{z}_0^{(k)}$.
  Decode $\hat{y}^{(k)} = D_R(\hat{z}_0^{(k)})$ and estimate $RR^{(k)}$.
end for
Compute $\bar{y} = \frac{1}{K}\sum_{k=1}^K \hat{y}^{(k)}$ and $u_{RR} = \mathrm{Var}(RR^{(1)}, \ldots, RR^{(K)})$.
Output: Posterior mean waveform $\bar{y}$, RR estimate, and uncertainty u_{RR}.

Table 2. Full respiratory-rate estimation results across datasets. MAE and RMSE are reported in breaths per minute.

Method	In-house RGB			AIR-125			Sleep		
	MAE↓	RMSE↓	r ↑	MAE↓	RMSE↓	r ↑	MAE↓	RMSE↓	r ↑
Peak/FFT	4.6	7.0	0.56	6.1	8.2	0.49	5.4	7.4	0.53
E2RespUNet	3.8	5.9	0.66	4.9	6.7	0.62	4.1	5.9	0.68
RespFormer	3.3	5.3	0.72	4.8	6.4	0.65	4.5	6.1	0.66
Latent regression	3.6	5.6	0.69	4.4	6.1	0.67	4.2	5.8	0.68
Waveform-space diffusion	3.2	5.1	0.73	4.7	6.3	0.66	4.1	5.7	0.69
ViTaL-Diff	**2.7**	**4.4**	**0.79**	**4.2**	**5.7**	**0.73**	**3.5**	**5.0**	**0.76**

A.7. Full Respiratory-Rate Results

The main paper reports MAE due to space limitations. Table 2 provides the full respiratory-rate estimation results using MAE, RMSE, and Pearson correlation r. MAE and RMSE are reported in breaths per minute. Lower MAE/RMSE and higher r indicate better performance.

ViTaL-Diff achieves the best performance across all metrics and datasets. On the in-house RGB dataset, it reduces MAE from 3.3 to 2.7 bpm compared with RespFormer and improves correlation from 0.72 to 0.79. On AIR-125, ViTaL-Diff obtains the lowest MAE and RMSE despite the dataset containing heterogeneous pediatric recordings with varied resolutions, frame rates, and visual conditions. On the Sleep dataset, ViTaL-Diff achieves the largest absolute gain over RespFormer, reducing MAE from 4.5 to 3.5 bpm and improving RMSE from 6.1 to 5.0 bpm. These results indicate that the latent diffusion formulation reduces both average error and large-error failures while better preserving agreement with the reference respiratory trend.

Compared with waveform-space diffusion, ViTaL-Diff consistently improves MAE, RMSE, and correlation. This supports the benefit of performing diffusion in a learned respiratory latent space rather than directly denoising raw waveform samples. Compared with latent regression, ViTaL-Diff also improves performance across datasets, showing that the gain is not only from tokenizing the respiratory waveform, but also from modeling a conditional distribution over plausible respiratory trajectories.

A.8. Evaluation Metrics

We evaluate respiratory-rate estimation using mean absolute error (MAE), root mean square error (RMSE), and Pearson correlation r. Given ground-truth respiratory rates RR_i and estimates $\widehat{RR}_i$, the metrics are defined as

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N} |RR_i - \widehat{RR}_i|, \tag{26}$$

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N} (RR_i - \widehat{RR}_i)^2}, \tag{27}$$

and

$$r = \frac{\sum_i (RR_i - \bar{RR})(\widehat{RR}_i - \bar{\widehat{RR}})}{\sqrt{\sum_i (RR_i - \bar{RR})^2}\sqrt{\sum_i (\widehat{RR}_i - \bar{\widehat{RR}})^2}}. \tag{28}$$

Lower MAE/RMSE and higher r indicate better agreement with the reference respiratory rate.

A.9. Uncertainty-Aware Evaluation

Table 3 reports the uncertainty-aware evaluation on the in-house RGB dataset. For each clip, ViTaL-Diff generates $K = 10$ respiratory waveform samples. RR is computed from each generated waveform, and the variance across sampled RR estimates is used as the uncertainty score u_{RR}. Clips with the highest uncertainty are rejected, and MAE is computed on the retained subset.

Table 3. Uncertainty-aware evaluation on the in-house RGB dataset.

Coverage	MAE↓	Relative reduction
100%	2.7	–
90%	2.3	14.8%
80%	2.0	25.9%
70%	1.8	33.3%

The monotonic decrease in MAE shows that u_{RR} is correlated with prediction difficulty. Rejecting the top 20% most uncertain clips reduces MAE from 2.7 to 2.0 bpm, while rejecting the top 30% reduces MAE to 1.8 bpm. This indicates that sample variability is not merely a byproduct of diffusion sampling, but a useful reliability signal. High-u_{RR} clips typically correspond to weak respiratory motion, occlusion, non-respiratory movement, poor visual quality, or ambiguous periodic patterns. The full uncertainty–coverage trend is visualized in Figure 5 (b).

A.10. Ablation Study

We include controlled ablations to analyze the contribution of each component. Table 4 summarizes the ablation results across datasets. Latent regression removes diffusion sampling and directly predicts respiratory latent tokens from video evidence. Waveform-space diffusion applies diffusion directly to respiratory waveforms instead of the learned latent space. Removing evidence dropout tests whether robustness comes from training with incomplete visual evidence.

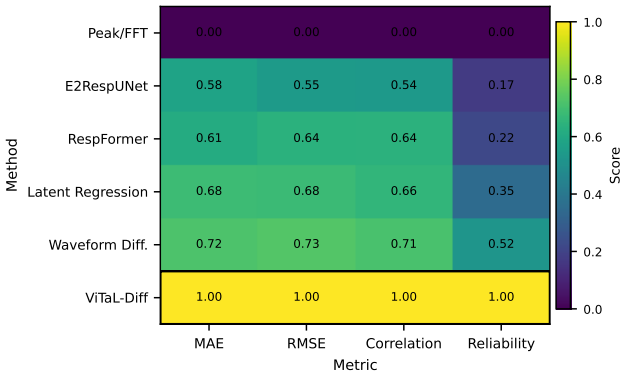
The ablation results show that each component contributes to performance. Removing latent diffusion causes the largest degradation on the in-house and Sleep datasets, indicating that deterministic latent prediction is less effective when visual evidence is ambiguous. Removing the respiratory latent space also increases error, especially on AIR-125, supporting the benefit of denoising structured latent tokens instead of raw waveform samples. Evidence dropout improves robustness across all datasets, suggesting that training with incomplete visual evidence helps the model handle occlusion, dropped frames, and weak motion. Removing uncertainty sampling does not change the deterministic architecture alone, but removes the posterior averaging and reliability signal, leading to worse MAE and eliminating the ability to detect unreliable clips.

Table 4. Component ablation across datasets. Each removed module increases MAE, highlighting the contribution of the full ViTaL-Diff design.

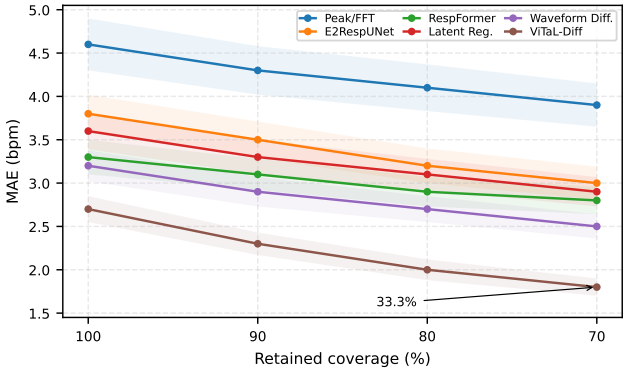
Variant	Dataset		
	In-house RGB	AIR-125	Sleep
Full ViTaL-Diff	**2.7**	**4.2**	**3.5**
w/o latent diffusion	3.6	4.4	4.2
w/o respiratory latent space	3.2	4.7	4.1
w/o evidence dropout	3.0	4.5	3.9
w/o uncertainty sampling	3.1	4.6	4.0

A.11. Extended Evaluation Analysis

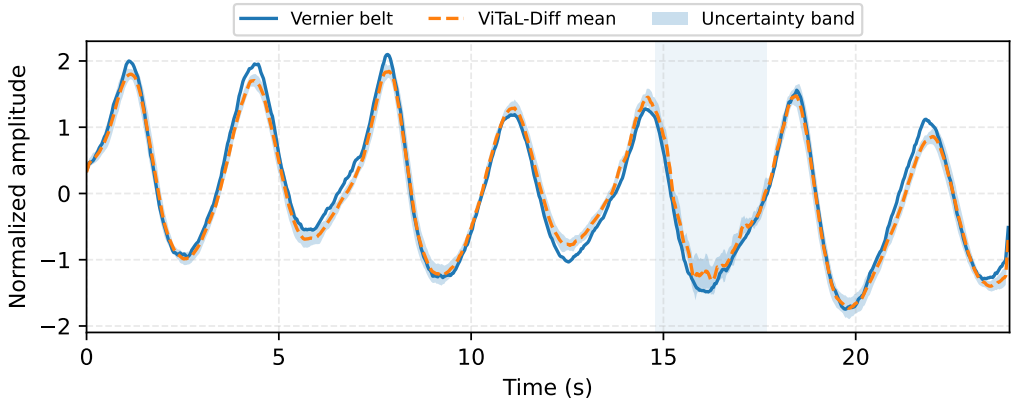
Figure 5 provides three supplementary analyses. Figure 5(a) summarizes normalized MAE, RMSE, correlation, and reliability scores, with all metrics scaled so that higher values indicate better performance. ViTaL-Diff achieves the strongest overall profile, consistent with the full quantitative results in Table 2. Figure 5(b) shows the uncertainty–coverage trend for all baselines. ViTaL-Diff maintains the lowest MAE across coverage levels, and its error decreases as high-u_{RR} clips are rejected, supporting sampled RR variance as a reliability measure. Figure 5(c) shows qualitative waveform reconstruction using the Vernier belt reference, ViTaL-Diff posterior mean, and uncertainty band. For this visualization, we use $K = 10$ generated waveforms per clip to compute the posterior mean and uncertainty band.



(a) Normalized multi-metric comparison



(b) Uncertainty–coverage trend



(c) Qualitative waveform reconstruction

Figure 5. Extended quantitative and qualitative evaluation. (a) Normalized MAE, RMSE, correlation, and reliability scores, where higher is better. (b) MAE after rejecting high-uncertainty clips across retained coverage levels. (c) Vernier belt waveform, ViTaL-Diff posterior mean, and uncertainty band from $K = 10$ generated samples.