We would like to thank the reviewers for their extensive and constructive feedback and are glad they found our work a *strong contribution*. We have made an effort to answer all their comments and will update our paper accordingly.

R2 & R5: Why is a high frame rate important? The standard sampling rate for contact pulse and respiration sensors is between 100-256Hz. This is because derivative information is often computed from these waveforms. For example, for heart and respiration rate variability measurement 20 Hz is not sufficient because peak timings are too imprecise.
R2 & R3: How does the illumination and surface reflection help in understanding the resp or pulse waveforms? The volumetric changes of blood in the surface of the skin cause changes in light absorption (↑ volume of hemoglobin = ↑ light absorption). This in turns affects the amount of visible light reflected from the skin, this is the source of the photoplethysmogram. The mechanical force of blood pumping around the body also causes subtle motions and these are the source of the ballistocardiogram. These color and motion changes in the video help us extract the pulse signal and heart rate frequency. Respiration is captured mostly via motions of the torso as the subjects breathes in and out. But, the principle of respiratory sinus arrhythmia also means that the color changes of the skin pixels due to blood flow also contain some respiration information. Separating these motions and color changes of interest from specular reflections and other motion noise is the primary task of camera-based physiological measurement.

R1: 1) We believe the computational benefits offered by our approach are a helpful contribution to the research literature. Applying knowledge distillation and model pruning to Hybrid/3D-CAN are definitely interesting lines of investigation but are beyond the scope here. 2) The attention module learns higher weights for skin regions in the pulse task, and torso regions in the resp. task. This boosts the SNR in the motion branch because changes from other sources can be more effectively ignored. Fig. 3-B shows that the temporal shifting in the motion branch can introduce more signal but also more noise, therefore the attention module is even more important. 3) Hybrid-CAN uses a 2D attention branch with an averaged frame whereas 3D-CAN uses a 3D attention branch with all the frames. We have shown in Table 1 that Hybrid-CAN was able to achieve a MAE of 1.12 while 3D-CAN had a MAE of 1.18. We experimented using the middle frame and last frame as the input to the 2D attention branch. We found that the mean frame provided the best results. Since the 2D branch only requires a single frame as input (the average frame), the generated attention mask can be shared with all the frames (10 in our case) in the motion branch, helping significantly reduce computation with little impact on accuracy. 4) TS-CAN sometimes performs better than MTTS-CAN because we train two separate "dedicated" TS-CAN models, one for pulse and one for resp. However, we only train one multi-task (MTTS-CAN) model which has to share the weights for two tasks and is much more computationally efficient but is not always as accurate.

R2: 1) We agree with that our approach could have applications beyond telehealth monitoring during a pandemic. We wrote our introduction to resonate with the current situation, but we agree it would be good to reduce this emphasis and highlight more applications in other domains. 2) We will add a breakdown of the performance by participant in the supp. material. The st.dev. across subjects was (MTTS-CAN): HR MAE: 1.15 BPM (AFRL) and 4.82 BPM (MMSE) and respiration MAE: 1.62 breaths/min (AFRL), these combined with the violin and Bland-Altman plots will be helpful. We agree about evaluation on patient groups of interest (e.g., COVID and A.Fib.) but a full clinical validation was beyond the scope of this work. Generally speaking we expect that improvements in accuracy on healthy people will also help with performance on patients. 3) Vital sign measurement is still imperative even if the cardiopulmonary symptoms are severe. Scalable and accessible health monitoring has many applications and could help democratize care.

R3: 1) The spatial resolution of the videos does not need to be very high (658x492 for AFRL) and we actually downsample the face regions to 36x36 pixels, which is quite effective (Chen & McDuff, 2018). Studies have shown that off-the-shelf webcams are sufficient for physiological measurement. We would argue that high resolution cameras are not necessary. 2) The reason we proposed the use of multi-task learning and tensor shifting is to enable our network to run on edge devices in real-time while maintaining a sufficient frame rate and state-of-the-art accuracy. Our goal in Sec. 3.2 was to provide an explanation of how to realize on-device real-time temporal and spatial modeling. Table 1 and Figure 3-A capture these results. 3) The Firefly RK-3399 has a CPU with 2 Cortex A72 (1.8Ghz) and 4 Cortex-A53 cores. This is equivalent to the CPU in the Snapdragon (Snap) 650 Mobile Platform[1]. Snap 650 is a mid-to-low-end mobile platform that Qualcomm launched in 2015. Snap 650 achieves a score of 275 in single-core evaluation and a score of 830 in multi-score evaluation[2]. The modern Snap 865 achieves a score of 903 in single-core evaluation and a score of 3304 in multi-score evaluation. To contextualize this, Snap 650 has a little worse performance than the Apple A9 chip used in the iPhone 6s. Thus, we believe that the computing platform we tested on is a fair benchmark for the class of mobile devices we are targeting. 4) Clothing and facial hair do obscure the skin and can affect the measurement of the pulse signal, indeed a limitation of optical measurement via PPG. Skin type can also have an effect, we coded the skin type and gender of subjects in AFRL using the Fitzpatrick scale. The results (MTTS-CAN) are as follows (Skin Type-MAE): I-2.37 BPM, II-2.62 BPM, 1.77-BPM, IV-0.06 BPM. (Gender-MAE): Men-1.34 BPM, Women-3.92 BPM.

R5: We will emphasis why our method makes the deployment of camera-based vital measurement more tractable given that the lower computation budget requires less power consumption per inference, we appreciate that suggestion.

[1] https://www.qualcomm.com/products/snapdragon-650-mobile-platform [2] https://browser.geekbench.com/