We sincerely thank all reviewers and the meta-reviewer for recognizing the contribution of our work and providing their valuable comments. We have modified our paper based on their suggestions and hereby list our improvements as follows.

Q1 (Reviewer 4)

The "human-centric" video is beyond human face images. The very-low bitrate coding capability of the method mainly relies on the highly structural characteristics of human face, while its effectiveness is questionable on other types of human-centric images.

A1 We have modified our "Introduction" section and clearly stated that we only focus on face-centric videos in this work. We have also added an additional Section 4.2.4 to list our limitations.

As pointed out by the reviewer, this paper shows the capability of combining sparse neural representation with high signal compression efficiency of existing coding methods. This approach can potentially be applied to other types of video content with highly structured features such as human body where a generic sparse neural representation can be learned. More detailed discussions can be found in Section 4.2.4.

Q2 (Reviewer 2, Reviewer 4)

The method is for single image compression. How are images like FFHQ put into video encoder?

A2 We have added more details in the "Experiments" section to clarify and explain the image-based evaluation. And we have added more discussions in section 4.2.4 to describe how to apply to video.

Our evaluation is similar to the Video Coding for Machine (VCM) standard (ref[17] in paper) where each frame is compressed individually as images. For VVC, the all-intra mode is used and each image is compressed with intra-prediction only. There are two main reasons the community takes the image-based evaluation. The first is the lack of large-scale labeled video datasets for recognition tasks (in our case for the face recognition task). The second is the lack of video-oriented models trained for such tasks. Since the focus of VCM and our paper is to develop compression methods for machine tasks, but not to restrict and retrain machine task models, image-based evaluation is generally used to be plugged into the existing machine task models.

Q3 (Reviewer 1, Reviewer 4)

Bitrates in tables and figure. Reviewer 4: The extra bit cost of the "Generic branch" and the "Domain-Adaptive branch" should be incorporated. Reviewer 1: bitrates are not equal, better to use QPs.

A3 We have modified the corresponding descriptions in the "Experiments" section to make it more clear how bitrates are computed.

To Reviewer 4: the bit costs in tables and figures include all extra bit costs from the "Generic branch" and the "Domain-adaptive branch" already. For our method, when used for human viewing, the bits include three parts: the generic $Y_{generic}$ and domain-adaptive $Y_{adaptive}$ for face region, and the VVC or LIC compressed bits for the remaining pixels. When used for machine analysis, the bits include two parts: the generic $Y_{generic}$ for face region, and the VVC or LIC compressed bits for the entire frame, which also contain $Y_{X_{lq}}$. In other words, when used for human viewing, $Y_{X_{lq}}$ is not used and therefore does not need to be transmitted and are removed from the overall bit count.

To Reviewer 1: It is a known challenge for NN-based compression methods to deliver exact bitrate match since there is no QP to tune, which is different from traditional coding schemes. As discussed in the "Experiments" section, we try our best to find the VVC configurations that can best match the very-low bitrates of our approach, and finally QP=30 and 42 are used for VVC.

Q4 (Reviewer 1, Reviewer 2)

Compare with deep video compression (Reviewer 1) or AV1/VP9 (Reviewer 2)

A4 Cheng2020 LIC (ref[6] in paper) and VVC are used as baseline methods to demonstrate the capability of our method in working with various codecs, NN-based or traditional. To the best of our knowledge, VVC has the SOTA performance better than deep video compression (DVC), and as pointed out by the reviewer, AV1/VP9 performs similarly to or lower than VVC. Cheng2020 gives SOTA performance on image compression and outperforms VVC as shown in JPEG-AI (ref[11] in paper). In other words, we try to use the SOTA compression methods in the field as baselines.

As discussed in Section 4.2.3, our work can also be seen as an approach to improve existing methods like VVC or Cheng2020 LIC, by using the sparse neural representation. Instead of competing with and replacing them, we combine the power of neural representation learning with the efficiency of existing baseline compression methods. It is within expectation that this combination can also improve other baselines like DVC, AV1/VP9.

Therefore, due to the space and time limits, we will report extensive evaluation of pairing with many other traditional and deep NN-based video coding baselines in the future extended version of the paper.

Q5 (Reviewer 1, Reviewer 4)

Time complexity. Online test adaption is time-consuming, especially for a large downstream task model, such as Swin-Transformer Big model.

A5 Our online test adaptation does not involve the large downstream task model. As a compression method we do not assume the availability of the end task model to tune our method. Instead, we use the online distortion loss through the facenet embedding, and the tuning is completely on the encoder side. The decoder time remains the same as the regular reconstruction. In addition, the online adaptation only involves multiple inference and gradient computation over the CFT_{code}/CFT_{lq} and reconstruction module. There is only one inference pass through the main embedding module and transformer.

We have added more discussions to make this more clear in Section 3.4.

Q6 (Reviewer 1, Reviewer 5)

Improve formats, captions, fonts, typos

A6 We have seriously proofread our paper and corrected the typos the best we could. We have also improved our figures and added more descriptions in the captions of the figures for better understanding.

Q7 (Reviewer 5)

Other quality metrics other than PSNR and SSIM

A7 We also used LPIPS as a metric for perceptual quality. We completely agree with the reviewer that PSNR/SSIM are not good for measuring the visual effect of the restoration tasks. As discussed in our experiments, our reconstruction looks much more visually pleasing than the VVC compressed result, but the PSNR/SSIM does not reflect this large difference. One of our future directions is to investigate on good perceptual quality metrics suitable for restoration tasks.