We thank the reviewers for their fruitful comments!

Response to Reviewer 2: We predict characters for Librispeech/Libri-light. Thank you for the pointer!

"when the official LibriSpeech LM ... is incorporated into decoding, it is not clear whether the experiments still represent a realistic low-resource scenario." - Please see Appendix C for results without any language model. For the main paper, we show results with the entire LM data similar to other recent work on low resource ASR (e.g., Park et al., 2020).

Response to Reviewer 3: Thank you for the pointers to related work, we will consider discussing them in the next version of the paper. We will also try to make it more self-contained given the space restrictions.

"I'm not convinced that this training works well conceptually." - The joint training poses challenges which we successfully tackled. It enables the model to learn units that are useful to solve the contrastive task instead of having to work with fixed units that may not be optimal. Moreover, in computer vision, similar approaches are achieving the best results for pre-training ("Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. Caron et al., 2020).

"... for ASR, we have a lot of transcribed data, and we can make a strong ASR model and perform transfer learning." - We respectfully disagree: while this statement is true for English and possibly a few other languages, the vast majority of languages have very little if any transcribed data. Moreover, our approach outperforms the best supervised models, even when training on all available transcribed data of Librispeech.

"... how to extract K detractors." - The distractors are quantized latent speech representations sampled from masked time-steps. If another masked time-step uses the same quantized latent, then it won't be sampled.

Response to Reviewer 4: We respectfully disagree that many design choices have not been justified: we provide a through evaluation of why quantizing targets is a good choice (see Sec 5.4). We compare our joint approach (quantization and context representation learning) to a pipelined approach (Discrete BERT). We motivate the diversity loss and why the encoder needs to be stabilized. Many other design choices such as gumbel softmax for quantization and the encoder network architecture have been evaluated in previous work.

"The paper would have been significantly different in terms of quality had you applied you approach to some standard semi-supervised learning tasks." - We focused on one of the largest datasets in the speech recognition community to demonstrate the effectiveness of the approach when large amounts of labeled as well as unlabeled data is available. This follows other recent work on semi-supervised methods for speech such as "Improved Noisy Student Training for Automatic Speech Recognition. Park et al., 2020" and "End-to-end ASR: from Supervised to Semi-Supervised Learning with Modern Architectures. Synnaeve et al., 2020" which achieve some of the strongest results.

Response to Reviewer 5: *"... from this work alone it's not clear why the proposed changes should work well for the problem domain. Moreover, why the interaction of the two proposed changes is so beneficial."* - The major changes to prior work are: (1) quantize only targets, (2) jointly learn the quantization and the contextualized representations as well scaling the model and data. (1) is thoroughly ablated in Sec 5.3. Our intuition is that discretized representations are more robust to artifacts in the data that generalize less well. This is useful for the learning task in the loss function (targets) but not when we want to build rich context representations (inputs). (2) is the major change to [1] "Discrete BERT" which is outperformed by our joint approach (see Table 1). This shows that joint training works better than pipelining discretization and context representation learning. The former enables adjusting the discretization when needed while the latter has to work with a fixed discretization which is less flexible.

Overall, our design choices are highly effective: wav2vec 2.0 outperforms the best other semi-supervised methods by a large margin on 100h labeled data and shows comparable results to the state of the art on 960h labeled data.

"Experiment 1: Why is Discrete BERT the only baseline that is evaluated in the limited regimes?" For the very low resource setups (10min, 1h, 10h), this is the only competitive baseline, the only other model is reported in the original Libri-light paper with far higher WER than Discrete BERT or our model (WER 92% (10min), 64% (1h), 44% (10h)).

"Experiment 2: Why are the methods featured in "Experiment 1" not also all included in this experiment?" - I believe you are referring to the 960h labeled data setup. Previous work simply did not report results for this high resource setup.

"It's not necessarily clear if this method is successful for its two-phase training regime. This method could trivially be extended so that it could iteratively apply it's two stages. I would be curious if this further improved performance."

Pre-training followed by fine-tuning is not an innovation of this paper. It has been previously applied to ASR in "Effectiveness of self-supervised pre-training for speech recognition". Baevski et al., 2020. Once the model is fine-tuned (second stage), pre-training on unlabeled data again is unlikely to benefit the model. Note that the first stage does not use any labeled data.