For the knowledge distillation setup, owing to the extra supervision by the feature level loss introduced from the pre-trained teacher encoder, the student encoder can be trained to approximate the behavior of the highly non-linear semantic space learned by a large amount of parameters in the teacher model, and then the representation from the student encoder can be used by the decoder shared from the teacher model to produce satisfactory results. However, for the teacher model which can only be trained by the reconstruction loss, such a highly non-linear representation space can only be modelled with encoder consisting of a large number of parameters. See the supplementary material for a cleared description on the knowledge distillation setup.