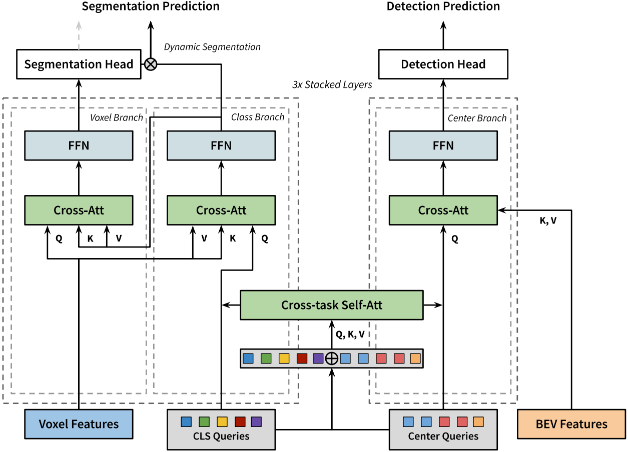

Figure 4: **Cross-task Transformer (XTF).** The segmentation and detection decoders share a self-attention layer to transfer the cross-task features. In the segmentation decoder, we use a bidirectional cross-attention to refine voxel features based on the aggregated class feature embedding. For simplicity, the skip connection and the layer norm are ignored in this figure.