We conducted additional training and testing on the septuplet dataset from Vimeo-90K, which is widely used for training video encoders. This dataset is generated from 89,800 video clips covering a wide range of scenes and various motion patterns. The resolution is 448x256, with a total of 91,701 samples.

We remove the sinusoidal embedding from CMA module and the down-sample operation before concatenation. There is no overlap between the videos used in the training and testing. Implementation details remain consistent with those in Sec. 4.1.

Experimental results:

$$Q_F = 56.91\%$$
$$\Delta \mathrm{R} = 2.15\%$$
$$T_{RC} = 0.63$$