

We conducted additional training and testing on the septuplet dataset from Vimeo-90K, which is widely used for training video codecs. 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.

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

Experimental results:

- 1) **Rate control accuracy and efficiency:** $\Delta R = 2.15\%$ $T_{RC} = 0.63$
- 2) **Quality fluctuations:** $Q_F = 56.91\%$