***Data sources*：**

The videos used in this research were sourced from math classes at a middle school in Zhejiang Province. We recorded the math classes of two groups over the course of a semester, using front and back camera angles. The videos captured both teacher and student interactions, with each class lasting 40 minutes. These classes were conducted by the same teacher who had been teaching for approximately five years, and were held in regular multimedia classrooms. The content covered the first semester of grade eight mathematics, with a focus on exercises. Both teachers and students granted their approval for the filming process..

***Data processing*：**

Data processing mainly includes concentration rating, screening of different concentration segments, video speech-text conversion, and dialogue text coding. The specific methods are as follows.

（1）Concentration recognition and scoring

In this study, we utilized a class concentration recognition method based on computer vision technology to identify and score students' group concentration. This involved face recognition, expression recognition, head posture recognition, and class concentration calculation. Firstly, the faces of students in the videos were detected. Next, the ResMaskingNet emotion recognition model was employed to recognize the students' expressions, enabling us to calculate emotional concentration. Subsequently, we leveraged the HopeNet model to detect head posture and determine behavioral concentration state. Finally, the average concentration of the entire class was calculated using specific weights.

（2）Fragment screening with different concentration

To identify and record relevant video clips, we utilized a concentration per second value. We screened and recorded video segments that met the conditions of continuous three seconds or more. The rationale behind selecting three seconds was based on our observation of classroom recordings, which showed that the teacher usually spoke for about three seconds at a time. The specific screening scheme is outlined below:

Step 1: We calculated the average concentration value and the highest concentration value of each video.

Step 2: Segments with high concentration levels were identified as those with values within a certain interval for three consecutive seconds. The exact time was then recorded.

Step 3: Segments with general concentration levels were identified as those with concentration values in another interval for three consecutive seconds, and the time was also recorded. It's worth noting that the classification of high and general concentration in this study was carefully considered.

Based on our recognition results and video recordings, we observed that when the concentration level was low, especially below the average, there was often silence in the class. The teacher did not speak, or there were no teaching-related activities occurring. Therefore, we focused our analysis on fragments with concentration values higher than the average.

（3）Video speech-text conversion and text encoding

In this study, iFlytek's interface was used to convert speech to text in videos. The converted text was then manually checked and adjusted to improve accuracy. Based on the concentration segments of the two sections, the corresponding video segment's dialogue text was screened out and encoded by paired trained researchers in pedagogy.