The following three figures depict the four metrics we employed to supervise the training process on the Cora dataset. These metrics include the micro F1-score, macro F1-score, class imbalance ratio (where a higher value is preferable), and label standard deviation (where a lower value is preferable). The results of four trials are recorded. It is observed that the curves of these metrics either consistently rise or fall at the onset of the training process and tend to converge within 1000 epochs. The stable and efficient learning process is attributable to two primary factors.

Firstly, the reward function we designed is goal-oriented, taking into account both the classification performance gain and the promotion of class diversity. Moreover, criteria-similarity and class-diversity metrics are involved in the state space, which provides further guidance for the agent to comprehend the relationship between the action chosen at the current state (which node to annotate) and the instant reward received. This is due to the reflection of class diversity information within both the state space and the reward function.

Secondly, the utilization of the advantage function as a baseline diminishes the variance of policy gradient estimation, allowing the A2C algorithm to update policies with greater stability. The guidance based on advantage evaluation has almost become a standard practice for policy gradient algorithms.