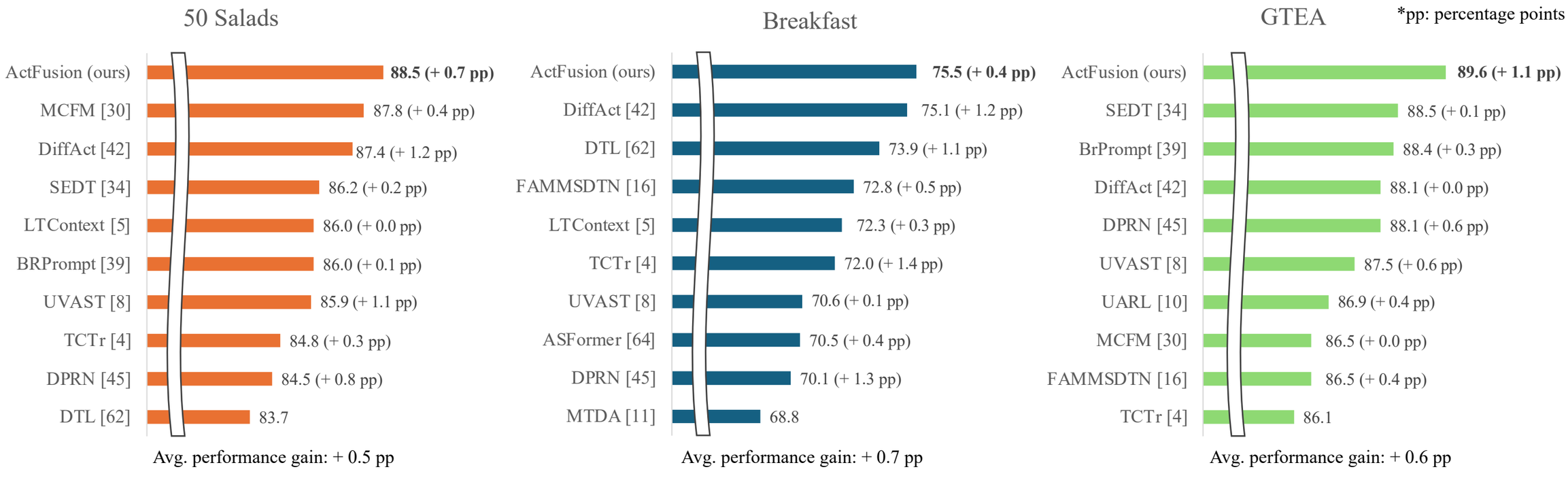

Figure R1. **Performance comparison of Top 10 models in TAS.** We illustrates the performance of the Top 10 TAS models for each dataset listed in Table 1, based on their average performance across all metrics. Those values inside the bracket indicates the performance difference between the adjacent models. The average performance gains are 0.5 percentage points (pp), 0.7 pp, and 0.6 pp for the 50 Salads, Breakfast, and GTEA dataset. ActFusion achieves 0.7pp, 0.4pp and 1.1 pp compared to the second-best model for each dataset, and 1.1 pp, 0.4 pp, and 1.2 pp compared to DiffAct, showing that the performance gains are meaningful at least on the two datasets.