Individual Task Completion Time Vs.  $F_1^{\mathcal{O}^*}$  Score Across Models 500 Model Random GPV-1 **BLIP** GPT4V **Ground Truth** 0.75 2 100 0.00.1 0.2 0.3 0.4 0.6 0.7 8.0 0.5 0.9 1.0  $F_1^{\mathcal{O}^*}$  Score