Response

I. RQ

Motivation&Approach. The goal is to evaluate separate weights of on-screen and cross-screen components weights on the repairing effects, we selected a representative application from our dataset, i.e., Google Translate, to conduct an experiment on a total of 20 test cases. Let α_1 and α_2 be the weighting value for on-screen and cross-screen components respectively. In this experiment, SEMTROID is used to repair test breakages with four combinations of (α_1, α_2), i.e., (0, 0.6), (0.3, 0.6), (0.6, 0), and (0.6, 0.6).

Table 1 shows the result. When selecting the same weighting coefficients, i.e., 0.6, the highest repair ratio (85%) is achieved. However, when only the cross-screen component was considered ($\alpha_1 = 0$, $\alpha_1 = 0.6$), the repair ratio dropped significantly to 65%, with only 13 test cases fixed. The performance under the (0.3, 0.6) and (0.6, 0) settings reached 80% and 75%, respectively. These findings suggest that incorporating both types of contextual semantics leads to superior repair effectiveness, with balanced weighting being particularly beneficial. Nonetheless, this is a preliminary experiment limited to a single application, and further large-scale empirical studies are required to validate the generality of these observations.

TABLE I
REPAIR EFFECTIVENESS WITH DIFFERENT α_1 AND α_2 ($\theta = 5$)

#breakages	**(0.6,0))**	**(0, 0.6))**	**(0.3,0.6))**	**(0.6,0.6))**
#breakages	15	13	16	17
#ratio	75%	65%	80%	85%

Different weighting values for on-screen and cross-screen contextual semantics will lead to different repairing effects. When selecting the same weighting coefficients for , i.e., 0.6, the best repairing effect is obtained.