Response

A. R1Q2

"Can the author conduct one ablation study that only adopts on-screen contextual semantics and another one that only adopts cross-screen contextual semantics?"

Motivation&Approach. The goal is to conduct an ablation study to evaluate the contribution of each semantics to the repair effects. We selected a representative application from our dataset, i.e., Google Translate, to conduct an experiment on a total of 20 test cases. In this experiment, SEMTROID is used to repair test breakages with only adopting on-screen contextual semantics and only adopting cross-screen contextual semantics.

Results. Table I presents the number of correct repairs made by SEMTROID only adopting on-screen contextual semantics (Sem+OS), only adopting cross-screen contextual semantics (Sem+CS), and adopting both (Sem+OS+CS). The number of total test breakages in this experiment is 20. Sem+OS+CS repairs the most breakages, then Sem+OS, finally Sem+CS. The results show that the two types of semantics contribute differently to test repairs. The repairing effect with combining on-screen and cross-screen contextual semantics can achieve the best repair ratio.

TABLE I
REPAIR EFFECTIVENESS OF ABLATION STUDY (α=0.6, $\theta = 5$)

	Sem+OS	**Sem+CS**	**Sem+OS+CS**
#breakages	15	13	17
#repair ratio	75%	65%	85%

Answer to R1Q2. The repairing effect with combining on-screen and cross-screen contextual semantics is superior to that of only adopting either one.

B. R3Q1

"It is recommended that the authors assign separate weights to on-screen and cross-screen components and conduct additional experiments with varying configurations to determine the most effective weighting scheme."

Motivation&Approach. The goal of this experiment 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). (We initially planned to conduct the experiment on (0.3, 0.6), but later we have not finished it due to time constraints)

Results. Table II shows the result. The number of total test breakages in this experiment is 20. When selecting the same weighting coefficients, i.e., 0.6, the highest repair ratio (85%) is achieved. However, when only the cross-screen component is considered ($\alpha_1 = 0$, $\alpha_1 = 0.6$), the repair ratio drops significantly to 65%, whereas the repair ratio under (0.6, 0) (namely only the on-screen component is considered) is 75%. The repair ratio under (0.3, 0.6) is increased to 80%. These findings demonstrate that incorporating both types of contextual semantics leads to superior repair effectiveness, with balanced weighting being particularly beneficial. However, 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 II
REPAIR EFFECTIVENESS WITH DIFFERENT α_1 AND α_2 ($\theta = 5$)

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

Answer to R3Q1. Different weighting values for on-screen and cross-screen contextual semantics will lead to different repairing effects. When selecting the same weighting coefficients for both types of semantics, i.e., 0.6, the best repairing effect is obtained.