We thank all the reviewers for their valuable comments. We will address the reviewers' concerns point by point.

Meta reviewers:
1) Regarding the concern of "no visual evidence", we have done the analysis of "qualitative results" and we believe that Figure6,7,8 show the effectiveness of the proposed method.
2) Regarding the concern of "introduction too general as there is only one task", we have changed the title to "DCMFNet: Deep Cross-Modal Fusion Network for Referring.
Image Segmentation with Iterative Gated Fusion" and emphasized the RIS task in the abstract. We reviewed the introduction and make sure that the RIS task is the main topic of the introduction section.
3) Regarding the concern about "lack of improvement over (and citation of) state-of-the-art results", we argue that the methods listed by reviewer R2 introduce semantic features that we do not use, and use pre-trained language models to encode these features, resulting in richer semantic embeddings, and therefore these methods achieve better performance than our proposed methods. It would be unfair to compare these methods with our proposed method. In fact, we mentioned that this could be our future work in the "limitation" of section 4.

Important points from reviewer R1:

1) Regarding the question "There are a lot of duplicates in Figures 1 and 3", we modified the content of both images and removed the duplicates in them.
2) Regarding the questions "The fusion unit structure image is missing" and "The fusion unit section is missing a corresponding icon", we have added the fusion unit structure image to Figure 4.
3) Regarding the question "In Table 1, it would be better if the best performers were bolded", we have changed the data in Table 1 as requested.
4) Regarding the question "The models corresponding to the three different network structures (resnet, transformer, darknet) in Table 1 should be compared separately", we have rearranged the results of the experiment as requested. The detailed results are shown in Table 1 of the paper.

Important points Reviewer R2:
1) Regarding the question "The results are not state of the art. For example, on the G-Ref dataset, DCMFNet-Trans (Ours) achieved 57.79, which is not as competitive as SOTA methods such as CRIS, CM-MaskSD, LAVT, JMCELN, VG-LAW, etc. And these methods are also not cited in the paper", we addressed this in the Meta Review section and cited all the references in "Limitations" of section 4.
2) Regarding the question "Lead feature and guided feature in Fig. 3 are quite confusing", we have changed Fig. 3. The detailed results are shown in Figure 3 in the paper.
3) Regarding the question "If the authors do not intend to test on different tasks, I suggest making the title more specific by mentioning the task of reference image

segmentation", we accept the reviewer's suggestion and changed the title to "DCMFNet: Deep Cross-Modal Fusion Network for Referring
Image Segmentation with Iterative Gated Fusion".