Figure 1: Qualitative results of gradient correction, the right column shows the zoom up area in left figure.

We appreciate all reviewers for the concrete and constructive comments. Following are our response for the concerns.

[Q1](R1, R2, R3, R4) The results reported in original paper of STM is much better than the implemented baseline and should be included in main results.

[A1] We insists on our reimplemented version of STM as baseline in our paper for following reasons. (1) The original STM did not open source training code and we find it hard to reproduce the results reported in the paper even when the training settings and data are aligned with the original paper. (2) The vanilla STM achieves good results mainly due to large amount of extra data (COCO, VOC, ECSSD etc.) for training, this results in unfair comparison with other methods since the data source are not the same. Therefore, we retrain our implemented STM model solely on training data of Youtube-VOS and DAVIS, which is the most common intersection of data source from previous works.

On the other hand, we agree that the original results in STM paper should be appended in main results for more complete comparison. Additionally, since vanilla STM open source the inference script and pre-trained model, we can combine it with our online gradient correction module in inference stage, we find this still brings around **1.9** boost in $\mathcal{J}\&\mathcal{F}$ score ($\mathcal{J}$ from 79.2 to **81.2** and $\mathcal{F}$ from 84.3 to **86.1**). This shows our method is general and can bring improvement for either the vanilla STM or our reimplemented version. These results and related discussion will be appended in updated paper.

[Q2](R1, R2) More visualization results on gradient correction module are required.

[A2] We visualize examples of the qualitative effect of gradient correction in Fig 1, we can see gradient correction can effectively suppress some false segmentation area and append segmentation of small part of objects.

[Q3](R2) The impact and motivation of gradient correction seems weak.

[A3] One core motivation of gradient correction is to mitigate the effect of low-quality reference mask during inference?. We emphasis that gradient correction can not only bring improvement for well-trained model, but also makes the model robust for partially low-quality reference sets. Please note the results of A.2 in Supplementary material Table 2, when some of the reference masks in the memory are replaced by rough bounding boxes or inaccurate prediction, gradient correction can effectively help the model yield high-quality prediction (at most $+7.3$ improvement over the perturbed model). Besides, it should be noticed that gradient correction is also helpful for the inference of vanilla STM (see **[A1]**).

[Q4](R2) The connection of three contributions seems weak.

[A4] The three contribution follows a naturally logical relation. The offline training scheme is first proposed and then it is extended to an complementary online version, and the cycle-ERF is derived from online update. Further, **all three contributions aim at better utilization of the information from previous reference masks.**

[Q5](R3, R4) Why not include extra data for training as other methods?

[A5] We agree including extra data source for training will boost the segmentation quality (e.g. **When MSRA10k is included, the performance on of our model on DAVIS17 validation set can be boosted from 71.7 to 76.2**). Nevertheless, the problem is there is no unified protocol constraining the allowed data source in the problem of video object segmentation for academical comparison. As a results, different method may rely on auxiliary data from different domain or tasks. This makes it difficult to conduct fair comparison since we can not align the data setting of all previous work. Therefore, we decide to report results of a purely trained model with DAVIS and Youtube-VOS data since they are the commonly used and necessary data source of previous works. We will also open source our training code and trained model, hoping this can be a more fair reference results of STM for future works.

[Q6](R4) The cycle-ERF is not clear.

[A6] The generation of cycle-ERF is nearly identical to that of gradient correction (Eq. 5) except for the cyclic reference set is an empty mask as "zero prior" and is gradually updated to show crucial area. In some case the cycle-ERF of a trained model highlights regions away from objects since it requires some contextual to better distinguish the objects under such "zero prior" condition.

Finally, we will carefully polish the writing and append related reference according to the suggestion from reviewers.