We thank reviewers for their comments. Our responses to reviewers' (R1-R4) comments are itemized as follows.

Q1. Key difference of (i) Adaptive Feature Bank (AFB) vs STM (R3), (ii) AFB vs other VOS work (R1, R4).

(i) 1) Our AFB is the first module that does not uniformly sample frames but dynamically manage object's key features in VOS literature. 2) STM uniformly stores every one of $K = 5$ frames in feature bank, which will fill up 1080Ti MEM (11GB) when a single-object video has 350+ frames. Practical videos are often longer (e.g., avg. YouTube video has ≈ 12 minutes or $22K$ frames). To process a 10-min video, STM needs to set $K = 300$ and misses many important frames (see **Q4**). In contrast, AFB performs dynamic feature merging and removal, and can handle videos with any length effectively. **(ii)** Most recent methods (other than STM) only store features from the first and/or last frames. Specifically, R4 mentioned four papers: DMM-Net [a] stores the 1st frame, FEELVOS [4] and RANet [6] store the first and latest frame. Motion-guided [b] is not a matching-based, but is in another category, mask-propagation based methods. Mask propagation methods are unstable when objects undergo significant occlusions. [4] and [6] were compared in Table 1. [a] has a $J\&F$ score of 70.7 on DAVIS17 (lower than STM 71.6, and ours 74.5). [b] reports J 76.4 and F 75.7 in its Table 4 on DAVIS16, while our scores are J 85.5 and F 83.4.

Q2. Key difference of Uncertainty Region Refinement (URR) vs existing work (R2, R4).

(R2) Thank you. We will discuss fine-grained segmentation work in **image segmentation** (e.g. *PointRend* [1] and *ShapeMask* [2]) in the revision. But our URR is the first module addresses boundary refinement in VOS literature. Meanwhile, the refinements in [1] and [2] are different from ours. *ShapeMask* [1] revised the decoder from FCN, and such global decoders can fail to recover accurate masks on boundaries. Our URR has a strong local regularization on uncertain regions to predict precise mask on boundary. *PointRend* [2] was published in CVPR2020, after this paper's submission. Its differences from ours are: **(i)** *PointRend* defines uncertainty on a binary mask, while ours is defined on more general multi-object classification (score ratio of the top-1 class to top-2 class). **(ii)** Furthermore, *PointRend* does a one-pass detection and refinement on uncertain regions. While we design an uncertainty loss to generate cleaner object masks (see Fig. 3). As R1 pointed out, this works like a binarization step, forcing the solution to be closer to the final/wanted 0/1 map. **(iii)** For refinement, *PointRend* refines uncertain points by their own features. While we use features from reliable points to refine the uncertain ones, through a non-local mechanism.

URR vs refinement module in [15, 22] (R4): Actually, the refinement module [15, 22] is used in our decoder to generate initial masks (see Line 97), which are inaccurate on boundary. Then, URR effectively refines these boundary regions.

Q3. Runtime performance analysis (R1, R3, R4): Our model has better runtime performance than the baseline STM. On DAVIS17, with an NV. 1080Ti, STM achieves 3.4fps ($J\&F$ 71.6), and ours achieves 3.9fps ($J\&F$ 74.5). Our runtime is a trade-off between latency and accuracy depends on requirement: if we limit the memory usage under 20%, it achieves 5.7fps ($J\&F$ 71.7). We will add detailed runtime comparison with other SOTA to the manuscript.

Q4. Why is AFB important? Is AFB w/o URR worse than baseline STM? (R2).

(i) AFB optimizes features management, leading to less memory usage (see **Q1**) and less runtime latency (see **Q3**) compared with STM. **(ii)** As Table 3 shows, compared with STM, AFB w/o URR has slightly downgraded $J\&F$ (from 72.2 to 70.2), but greatly reduces ($\sim 75\%$) memory usage. Current benchmarks YouTube-VOS (132 frames avg.) and DAVIS17 (67 frames avg.) only contain short clips, which cannot show the advantages of AFB in dealing with long videos in real-world (see **Q1**). **(iii)** A new dataset of long videos (2K+ frames each) is added to show the performance in real-world, on which AFB w/o URR vs STM are 82.9 vs 66.5 in $J\&F$ score. This dataset will be released with paper.

Q5. Effectiveness of two sub-components of URR: uncertainty loss (R2), and local refinement mechanism (R1).

We performed these two new experiments (b, c) on DAVIS17. The $J\&F$ scores are: (a) AFB only: 70.2; (b) AFB + uncertainty loss: 72.7; (c) AFB + fine-grained module: 72.1; (d) AFB + URR (full model): 74.5.

Q6. On DAVIS17: Qualitative comparison with STM (R2)? STM has better scores in Leaderboard (R3)?

(R2) We followed most SOTA settings that train models only using DAVIS17. The released STM checkpoint was trained on DAVIS17+YouTubeVOS. Also, STM authors didn't release training codes or corresponding results for qualitative comparison. (R3) The STM scores in leaderboard was also trained on DAVIS17+YouTubeVOS, we used the right score that only uses DAVIS17 from its paper.

Q7. Red Sign: On YouTubeVOS unseen, lack comparison by running SOTA (STM) in same environment (R2).

We can only compare numbers, not re-run STM codes because (a) STM didn't release training codes (see STM GitHub *Reproducibility* Issue #6); (b) STM didn't plan to release model for YouTubeVOS comparison (GitHub Issue #3).

Q8. Performance for only pretraining on image sets in ablation study (R1): On DAVIS17, only pretrained by image datasets: $J\&F = 60.9$; after training on DAVIS17, $J\&F = 74.5$.