Paper ID 8972 - Rebuttal

We thank the reviewers for their positive feedback. **R123:"strong results", "surpass the state-of-the-art", "evaluation backups the method's effectiveness"**. R1:"method technically solid", "evaluated on two different tasks"; R2:"paper is well-written and the method is easy to follow", R3:"Ablation studies are thorough", "impressed by the details and additional experiments".

R1:"components are not effective ... Tab. 4". For the simpler AVOS task, Table 3 ablations show our multiscale memory is improving 2-3% on three datasets. FS-VOS faces two additional challenges that have complex interaction with our contribution: (1) Confusing support examples (L557-559) & Zhang et al. CANet. CVPR 2019; (ii) Sole reliance on correlation tensors can degrade boundary precision (L848-850) & [11]. We demonstrate these interaction effects in rebuttal Table 1, also partially reported in Suppl. Table 6. Our method with adaptive k-shot or boundary refinement beats the baseline with these components by 1.2-2.3%, which confirms the benefit is not from the added components but rather from our multiscale memory decoder. Moreover, Suppl. Figs 2 & 4 show that working alone our core components improve robustness by up to 2%. Finally, we show in Suppl. (L583-624) using bigger memory helps in learning different parts of the novel class/background. UVOS *vs.* camouflaged object detection (COD) ... different tasks UVOS, which following [37] we call Automatic VOS (AVOS), is concerned with automatically segmenting the primary object in terms of saliency, which can be from appearance, motion or both. In COD, emphasis is on motion only and it is standard to consider it UVOS [37]. Comparing UVOS methods on a COD benchmark is conducted routinely [14] and Suppl.[36]. Learning the proposed method using the same training dataset? We train on YouTube-VOS and DAVIS, as do other AVOS methods (Suppl. L262-264). Our AVOS baseline also uses the same training setup as ours, yet we considerably outperform it.

Method	mIoU				
	Fold 1	Fold 2	Fold 3	Fold 4	Mean
Base-AK	49.1	70.8	**65.5**	65.4	62.7
Ours-AK	**52.8**	**72.4**	64.1	**66.1**	**63.9**
Base-BR	48.0	71.6	64.1	67.5	62.8
Ours-BR	**52.2**	**73.8**	**65.7**	**68.7**	**65.1**

Table 1: Adaptive k-shot (AK) & boundary refine (BR) ablation on multiscale baseline (Base) and Ours.

The compared VOS methods are not SOTA anymore. Four methods are cited, two of which were published after our submission. Table 2 has comparisons for the others. Our method is either on-par or outperforms them without the use of optical flow, while HFAN requires optical flow. We will add them to the final version. writings are hard to follow No other reviewer commented negatively on our presentation style, text or figures; indeed R2 says "well written and easy to follow".

Method	mIoU		
	DAVIS	MoCA	YTBO
Iterative	85.6	-	-
HFAN	**87.4**	59.9	73.4
Ours	86.7	**80.3**	**78.2**

Table 2: Iterative: Lee et. al., AAAI 2022. HFAN: ECCV Pei et. al. 2022.

R2: The motivation from FS-VOS to AVOS Our multiscale memory approach is applied flexibly to FS-VOS and AVOS. FS-VOS has two main challenges beyond AVOS: confusing support examples (L557-559); sole reliance on correlation tensors degrades boundary precision (L848-850). Notably, these challenges persist in single image Few-Shot Segmentation (FSS). AVOS is used as a simpler setup without these challenges to ablate our method. Moreover, the current FS-VOS task is under-explored with only one benchmark [3]. So, we use AVOS for additional evidence. Erroneous correlation at different scales ... references. This challenge previously has been established: HSNet [20] ablated shallow *vs.* deep *vs.* all layers in computing correlations. They showed that use of shallow or deep only, incurs worse performance, stemming from erroneous correlations. Motivation of the adaptive K-shot It is motivated by the need to reduce impact of confusing support examples (L557-559), also reported elsewhere (Zhang et al. CANet. CVPR 2019); Table 4 confirms the benefit. Ablation of the memory Table 5 does not include adaptive k-shot or boundary refine; improvements with more memory entries are small because memory has biggest impact in interaction with these other modules, due to FS-VOS challenges noted above & supported by Table 1 rebuttal. While it is beyond a rebuttal to do a detailed memory entry ablation, results for 5 entries on folds 1-4 are 51.3, 67.3, 63.6 & 64.6%, resp.

R3: Time dimension. our FS-VOS approach implicitly uses the temporal dimension, as it meta-learns multiscale memory entries through cross attention with temporal data, which drives the model to learn temporally consistent attention maps, as shown in the Suppl. video. We will clarify this in the final version. Is method compared to the 'right' competitors. Since FS-VOS is under researched, we already compare to the best alternative, DANet [3], on the standard benchmark; no others have been suggested in the review. Still, we provide additional comparisons, TTI (only published on arXiv), its single image FSS competitor, RePRI, and the more recent ASNet FSS method. ASNet is a recent development on HSNet[20] and we used the same training setup for both. Table 3 shows our method outperforms these by notable margins. Query clips ... only single object. Query clips can contain multiple objects: Suppl. video Ex. 1, 2 have two objects from different and same categories, resp.; Ex. 5 is a challenging scenario where the object to be segmented ('Hand') is not salient. Our method segments 'Hand' and differentiates it from other objects, while our baseline severely fails.

Method	1	2	3	4	Mean
DANet	43.2	65.0	62.0	61.8	58.0
RePRI	45.8	68.6	59.3	64.2	59.5
TTI	48.4	68.5	62.6	62.4	60.5
ASNet	48.0	70.0	64.0	66.9	62.2
Ours	**52.2**	**73.8**	**65.7**	**68.7**	**65.1**

Table 3: YouTube-VIS FS-VOS mIoU: 4 folds, 5-shot. DANet [3]; RePRI: Boudiaf et al. CVPR 2021; TTI: Siam et al. arXiv:2203.14308 2022; ASNet: Dahyun et al. CVPR 2022.