Rebuttal: Multiscale Memory Comparator Transformer for Few-Shot Video Segmentation

We thank reviewers for their positive comments: R1: Strong results across several settings; Well ablated design decisions R2: preserves the detailed feature maps and reduces confusion between the novel class and background; R3: The proposed technique is reasonable and solid. We next address each reviewer's individual questions; subsequently, we address questions from multiple reviewers.

R1: details of meta-learning Secs 4.1, 4.2 describe meta-learning (episodic training). Features are extracted from both support and query sets (L511), then correlation tensors are computed (L517) as input to multiscale memory decoding (L522-530). So, the memory entries are meta-trained on the support-to-query correlation tensor and change based on the query. During inference, changes in the query set will affect the correlation tensors and subsequently affect the final predictions; hence, the meta-learning scheme.

review...(multigrid) Classical methods cited, L146 [2]; will reference more with discussion in final version. No previous work has explored multigrid transformer decoding.

appearance bias We do well on camouflaged animals (MoCA) where motion is critical for segmentation, showing little appearance bias. A2D training data is appearance biased (segmentation typically doable from actor appearance alone), correspondingly biasing algorithms trained on it.

weak baseline Baselines are strong for 2 main reasons: 1. In both FS-VOS and AVOS tasks our baselines (coarse-to-fine) outperform state of the art with good margins: Tables 1,4 (FS-VOS); Tables 2,3 (AVOS); 2. Baselines are adaptations of recent work from top-tier venues: Mask2Former [6] (CVPR'22) for AVOS; HSNet for FS-VOS [18] (ICCV'21).

multi-grid...commonplace Multigrid computer vision was popular in the 1980's (*e.g.* Terzopoulos and others), but recently has seen less attention. No transformer decoders have considered multigrid. By definition, multigrid entails bidirectional information exchange between scales [2]; unidirectional (*e.g.* coarse-to-fine) technically is not multigrid. Another common multiscale approach atrous pyramid (COSNet[17], PMMs[35]), is in methods we outperform.

results...grouping by data We seek an approach that is not task/dataset specific. So, we organize by contribution.

R2: sensitivity to hyperparameter..specific architecture The only extra hyperparameter is number of memory entries, ablated in Table 5. Others are similar to baseline; see Suppl L228-251. Across multiple backbone architectures, we show improvement over baseline, Suppl Table4.

generalize to real-world MoCA is camouflaged animals in their natural settings; YouTube-VOS is in the wild YouTube video. Both are reasonably representative of real-world, *e.g.* complex backgrounds & lighting variation.

R3: components high-level figures. Description of our core contribution, $D_{multiscale}$, with diagram is within Fig. 2 & L376-379; $D_{hypercorr}$ defined, L518-521, & standard [18]; D_{refine} precisely defined in Eq. 8 & L458-467.

abstract We define problem (L16-17), challenge (L18-24) & our solution (L24-28). The rest has more details.

after reading the full paper, I get the idea of comparator. Comparator explicitly defined early on, L99-100.

contribution 2 and 3...duplicate Contribution 2 is multiscale memory transformer decoding to output dense feature maps instead of a compressed representation; contribution 3 is study of the best information exchange between these dense feature maps, *i.e.* stacked vs multigrid. They are assessed separately in Tables 3 & 4 showing nontrivial improvements per component across different datasets.

decoding needs to be defined We explicitly define it as transformer decoding, Fig1 caption (L88-89) & L123-130.

novelty w.r.t MViT, MViTv2, and MemMViT. We are 1^{st} to explore multiscale memory transformer decoding in meta-learning for few-shot learning & introduce multigrid bidirectional exchange in attention heads. MViT & MViTv2 use transformer encoding; they have neither a transformer decoder nor a memory module. MemMViT has a memory module, but only in the encoder & for a different purpose, extracting information across multiple clips in a video. Our multiscale memory is in the decoder for dense prediction, with purpose of maintaining spatiotemporal features without compressed output. Also, our memory module: 1) has entries that learn different parts of the novel class vs background from a single video, see Supp video; 2) multiscale structure learned in a meta-learning scheme that depends on support-to-target query relations; 3) exploits bidirectional multigrid information exchange. Previous work lack 1-3.

memory...used It is used in both training & inference; will clarify in final version. writing is hard to follow Our paper conveyed the main message properly, as confirmed by R1:MMC Transformer utilizes cross-attention between backbone features on target query and meta-learned memory features to R2:preserve the detailed feature maps and reduce confusion between the novel class and the background & neither state global writing concerns. We've responded to all specifics; will add all minor suggestions to final version.

R13: Figure 3 Will split into 2 figures in final version, as R3 suggests. Will elaborate caption for the bottom part, which will be its own figure, stating it shows multiscale memory decoding preserves output spatiotemporal dimension, $O \in \mathbb{R}^{TH_lW_l \times D}$, while query decoding compresses, $O \in \mathbb{R}^{N \times D}$, in response to clarification sought by R1.

R23: params, inference time, training memory Parameters for our full approach vs baseline 28.6M & 28.1M, resp., *i.e.* small increment. Inference time in Suppl L363-376, reducing FS-VOS state of the art 7x. Training memory in Suppl L477-529. Will bring all numbers into final version.