Many of the recent successful methods for video object segmentation (VOS) are overly complicated, heavily rely on fine-tuning on the first frame, and/or are slo