URVOS: Unified Referring Video Object Segmentation Network with a Large-Scale Benchmark

Supplementary Document

A Dataset

A.1 Dataset statistics

While A2D Sentences have 8 object categories (*e.g.* adult, baby, ball) and J-HMDB sentences only consider human, our Refer-Youtube-VOS has 94 categories including animals (*e.g.* eagle, snake, zebra), vehicles (*e.g.* truck, airplane, bus), sports (*e.g.* frisbee, skateboard, bike), electronics (*e.g.* camera, microphone, watch), common objects (*e.g.* cloth, tissue, shovel) and person. The distribution of objects per category in our dataset is illustrated in Figure 1, where it follows a long-tail distribution. The size of the vocabulary used for our dataset is 12,099. We also add "UNK" token to handle out-of-vocabulary words. Figure 2 visualize the frequency of words used in referring expressions. For the Full-video expression, each expression has 10.1 and 9.6 words on average in train and validation set, respectively. For the First-frame expression, it has 7.4 and 8.3 words on average.

A.2 Dataset collection

We give detailed instructions for the annotation and provide three annotation examples. For each example, we give both good and bad descriptions to help workers annotate correctly. Bad descriptions contain common errors, *e.g.*, too comprehensive to refer to the exact target object. Workers can remark "unknown" if the target object is hard to identify, and we drop the object if multiple "unknown"s are found in an object. Before collecting annotations, we conducted a validation test to select workers for the main task. We monitored the quality of results and work time, and selected about 50 workers who give concise and precise descriptions consistently. We then conducted verification steps with another 20 verified workers after finishing initial annotations. The workers evaluated the annotations of how well they localize the target objects quantitatively in a 1-to-3 scale. Each annotation was verified by 4 workers. If a target object fails to be localized by more than 1 worker, then we drop the annotation.

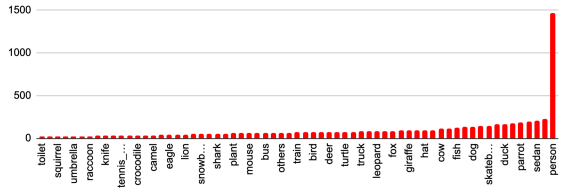


Fig. 1: The distributions of objects per category in Refer-Youtube-VOS dataset.

Fig. 2: The word cloud of referring expressions in Refer-Youtube-VOS dataset with the most frequent words.

B Additional Experiments

Annotation Type We vary two types of annotations on train and validation set, and conduct experiments with all combinations. As introduced in Section 3 of the main paper, full-video expression denotes that annotators are given the entire video for annotation, while first-frame expression denotes they only use the first frame of each video. The results are listed in Table 1. For the same validation set, the full-video expression gives superior results to the first-frame expression as the full-video one contains more relevant and richer information of video. However, the gap is not significant, which is partly because the videos in our dataset are 4-6 seconds long on average and the impact of two different annotation types is not salient. Using both full-video and first-frame expressions further improves the model performance.

Effects of feature levels We test different levels (Res4, Res5) of visual features for memory and cross-modal attention in our model. As shown in the Table 2, Res5 features for cross-modal attention and Res4 features for memory attention give the best performance.

Table 1: Ablation study on the annotation type of referring expressions.

Train	Validation	prec@0.5	prec@0.6	prec@0.7	prec@0.8	prec@0.9	$\mathcal{J}$	$\mathcal{F}$
Full-video	Full-video	51.02	45.03	38.63	28.70	14.06	44.18	48.75
First-frame	Full-video	49.04	43.76	37.30	28.43	13.58	42.37	46.75
Full-video	First-frame	51.49	45.67	38.66	29.33	14.20	44.70	48.76
First-frame	First-frame	51.04	45.14	38.37	28.98	14.24	44.11	47.85
Full-video	All	51.25	45.34	38.64	29.01	**14.13**	44.43	48.75
First-frame	All	50.02	44.43	37.83	28.70	13.90	43.23	47.29
All	All	**52.19**	**46.77**	**40.16**	**29.68**	14.11	**45.27**	**49.19**

Table 2: The Effects of feature levels for cross-modal and memory attention on Refer-Youtube-VOS dataset with Full-video expressions.

Cross-modal	Memory	$\mathcal{J}$	$\mathcal{F}$
Res4	Res4	42.98	47.17
Res5	Res5	43.90	47.97
Res5	**Res4**	**44.18**	**48.75**

C Additional Qualitative Results

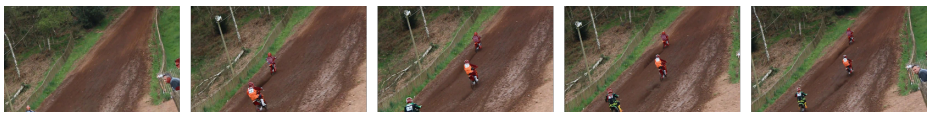
We present more qualitative results of our model on our Refer-Youtube-VOS dataset and the Refer-DAVIS$_{17}$ dataset in Figure 3 and 4, respectively. Our model segments out the exact target instance with sharp boundaries among multiple objects in the presence of some practical scenarios, *e.g.*, occlusions or shape deformation.

Failure Cases We also show some failure cases in Figure 5. These happen in challenging scenes, in which the target object is not salient or is hard to differentiate from other objects, requiring complex descriptions. Some of these cases can be handled by annotating more and richer language expressions for training.

(a) *A bus is moving in the road from right to left.*

(b) *A bike being used to do tricks.*

(c) *A kangaroo laying down behind another kangaroo.*



(d) *A person riding a motorbike in front of another wearing an orange helmet.*

(e) *An owl which is sitting on left eat something from its leg.*

Fig. 3: Additional qualitative results of our models on Refer-Youtube-VOS dataset.

(a) *A man jumping across fences.*

(b) *A man in a suit riding a scouter.*

(c) *A black colored pig.*

Fig. 4: Qualitative results of our models on Refer-DAVIS$_{17}$ dataset.

(a) *A person is walking on the road towards right holding a bag in her left hand.*

(b) *A lion is on the right side standing on a rock looking down at others.*

(c) *A tiger on its back, playing.*

Fig. 5: Some failure cases of our models on Refer-Youtube-VOS dataset.