Weakly-Supervised Referring Video Object Segmentation through Text Supervision

Supplementary Material

Overview of supplementary material

This supplementary material provides 1) details of datasets; 2) additional ablation studies on Refer-YouTube-VOS [37]; 3) more visualization results.

S1. Details of Datasets

A2D-Sentences [14] contains 3,754 videos, split into 3,017 for training and 737 for testing. Each video is annotated with three or five frames, providing pixel-wise segmentation masks for various target instances. In total, the dataset includes 6,655 referring expressions, each of which corresponds to an instance within the annotated frames. Refer-YouTube-VOS [37] comprises 3,978 videos annotated with 15,009 referring expressions. Pixel-wise segmentation masks are provided for every fifth frame in these videos. Furthermore, following [32], we evaluate the models trained on A2D-Sentences and Refer-YouTube-VOS by testing them on JHMDB-Sentences [14] and Refer-DAVIS17 [18] without finetuning, respectively. These two datasets extend the original vision-only datasets, JHMDB and DAVIS17, by incorporating rich text annotations. Specifically, JHMDB-Sentences is comprised of 928 videos, each paired with a referring expression. In contrast, Refer-DAVIS17 includes 90 videos and is annotated with a total of 1,544 referring expressions.

S2. Additional Ablation Studies

S2.1. Contrastive referring expression augmentation

Number of generated expressions. We vary the number of generated positive expressions P from 2 to 10 and that of generated negative expressions N from 24 to 72 on Refer-YouTube-VOS. The results on the validation set and test set are presented in Fig. S1. Based on the validation set, we select $P = 6$ and $N = 48$ as our default settings, and this configuration also achieves the best performance on the test set.

Threshold in CREA. We vary the cosine similarity threshold in CREA from 0.7 to 0.9 on Refer-YouTube-VOS. The results on the validation set are presented in Tab. S1. It demonstrates that our default setting (WSRVOS w/ CREA(0.8)) performs the best.

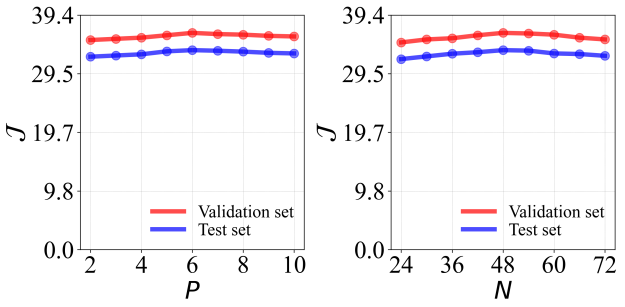

Figure S1. Parameter variation for the number of generated positive and negative expressions on the validation set and test set of Refer-YouTube-VOS, where the red line denotes the validation set and the blue line denotes the test set.

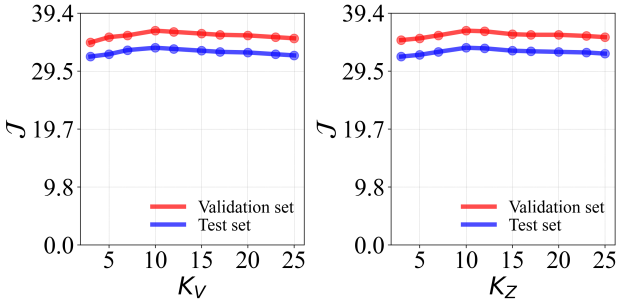

Figure S2. Parameter variation for the number of selected features in the bi-directional vision-language feature selection module on the validation set and test set of Refer-YouTube-VOS, where the red line denotes the validation set and the blue line denotes the test set.

Method	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
WSRVOS w/ CREA(0.7)	33.1	39.1	36.1
WSRVOS w/ CREA(0.9)	33.2	39.2	36.2
WSRVOS w/ CREA(0.8)	**33.5**	**39.4**	**36.5**

Table S1. Ablation study on the CREA module.

S2.2. Bi-directional vision-language feature selection

Position of the bi-directional vision-language feature selection module. Our proposed bi-directional vision-language feature selection module is integrated after the 3rd, 6th, 9th, and 12th layers of encoders. We also evaluate an alternative variant, denoted as V-L selection (All layers) in Tab. S2, where the module is inserted after every encoder layer. However, this variant does not lead to clear performance improvement while substantially increasing the computational cost (training time increases from 12.3 hours to 14.1 hours

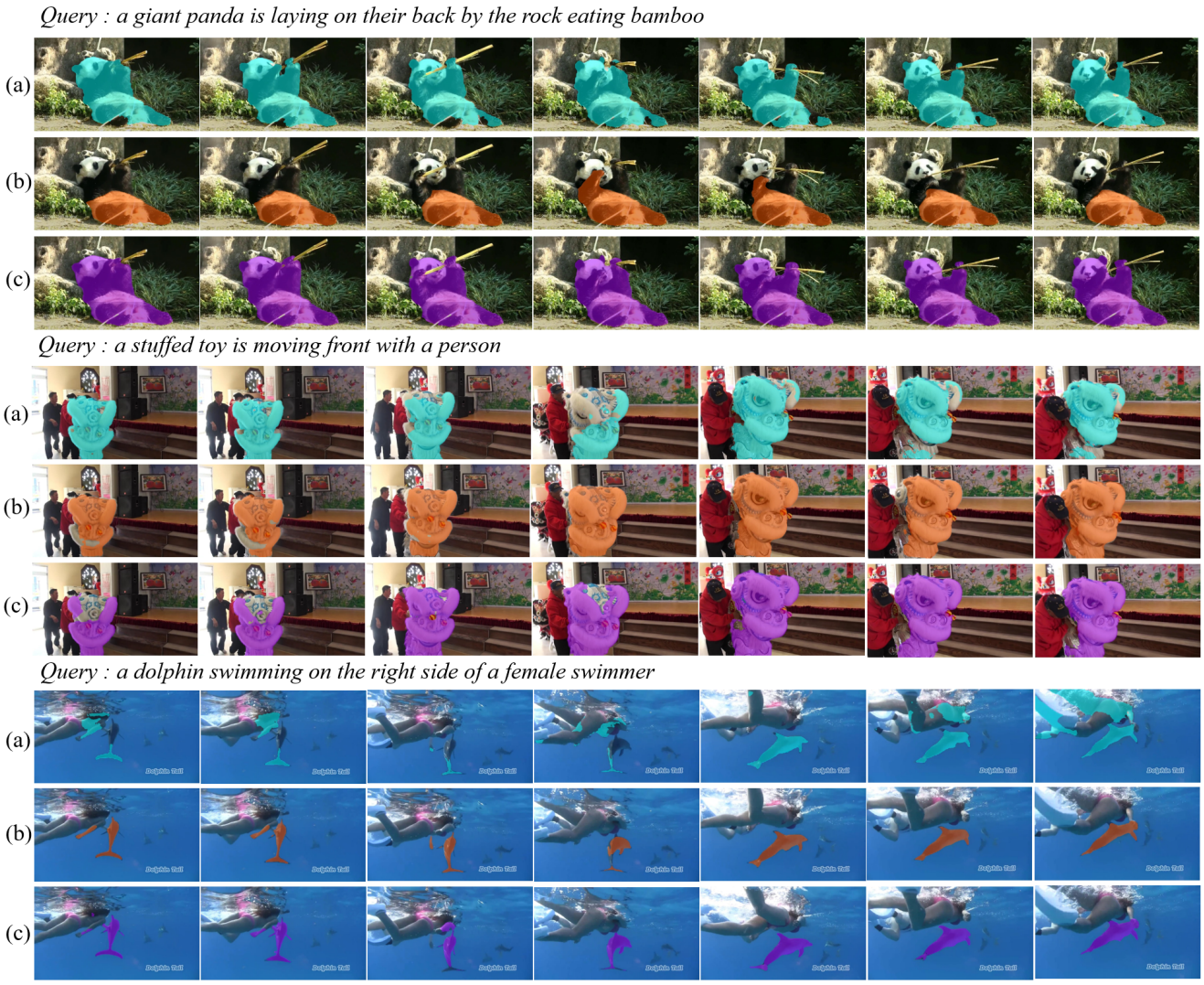

Figure S3. More visualization results on Ref-YouTube-VOS. (a) DViN (adapted weakly-supervised RIS method) [8], (b) OCPG (point-supervised RVOS method) [40], (c) our proposed WSRVOS.

and inference speed decreases from 58 FPS to 45 FPS).

Number of selected features. Fig. S2 presents the model performance under different numbers of selected visual and linguistic features, *i.e.* K_V and K_Z, on the validation set and the test set of Refer-YouTube-VOS. The results on the validation set indicates that $K_V = 10$ and $K_Z = 10$ achieve the best performance, and this setting consistently achieves the highest performance on the test set.

S2.3. Temporal segment ranking

In Tab. S3 we introduce a variant in which we replace the temporal segment ranking (TSR) constraint with a pairwise segment consistency (PSC) loss, *i.e.*, we simply encourage the IoU between the segmentation masks of each pair of

Method	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
WSRVOS w/o V-L selection	30.7	37.3	34.0
V-L selection (All layers)	33.4	**39.5**	**36.5**
WSRVOS	**33.5**	39.4	**36.5**

Table S2. Ablation study on the bi-directional vision-language feature selection module.

Method	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
WSRVOS w/o TSR	29.3	34.9	32.1
TSR → PSC	32.6	38.7	35.6
WSRVOS	**33.5**	**39.4**	**36.5**

Table S3. Ablation study on the temporal segment ranking constraint.

consecutive frames to be as close to 1. We can observe that this variant (TSR → PSC) outperforms the variant without TSR (WSRVOS w/o TSR), but still performs inferior to our original WSRVOS with TSR.

S2.4. Warm-up strategy

In Tab. S4 we introduce a variant in which we apply a warm-up strategy during the early training stage, *i.e*., only the classification loss is optimized in the first 5 epochs, while the remaining losses are activated afterward. We can observe that this variant (WSRVOS w/ warm-up strategy) yields no performance gains. Therefore, no warm-up strategy is employed by default.

Method	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$
WSRVOS w/ warm-up strategy	33.4	39.4	36.4
WSRVOS	**33.5**	**39.4**	**36.5**

Table S4. Ablation study on the warm-up strategy.

S3. More visualization results

We present additional qualitative results of WSRVOS and two comparison methods, *i.e*. DViN (adapted weakly-supervised RIS method) [8] and OCPG (point-supervised RVOS method) [40]), across video frames in Fig. S3.