Supplementary Materials for Learning Cross-Modal Affinity for Referring Video Object Segmentation Targeting Limited Samples

Guanghui Li[1*], Mingqi Gao[2,3*], Heng Liu[1†], Xiantong Zhen[4], Feng Zheng[2†]
[1] Anhui University of Technology, [2] Southern University of Science and Technology,
[3] University of Warwick, [4] United Imaging
guanghui.li1998@gmail.com, mingqi.gao@outlook.com, hengliusky@aliyun.com,
zhenxt@gmail.com, f.zheng@ieee.org

1. Additional Dataset Details

Mini-Ref-YouTube-VOS is a large Few-shot RVOS dataset containing 1668 videos, more than 2500 instances, and about 5K natural language descriptions. This dataset contains 48 categories, covering the most common categories in natural scenes, which can effectively simulate natural scenes in the real world.

Mini-Ref-SAIL-VOS is a dataset extracted from games, and the data is different from natural scenes. This dataset is somewhat challenging due to its data originating from game scenarios. Specifically, occlusion and multi-target phenomena exist in most videos.

In addition, since Ref-COCO or Ref-YouTube-VOS cover many categories, we choose categories the model has never encountered for testing. We have found several video sequences of common items in daily life. Notably, these video sequences do not have corresponding natural language descriptions, so we similarly add referring expressions to them.

2. More Visualization Results

We show more visualization results of the model. As shown in Figure 1, Figure 2, it can be seen that our model can still achieve excellent results even in the face of data from different scenes.

3. Ablation Study for Text information

To verify the role of text information, we tested several selected unseen category data in the experiment. Since object masks will become inaccurate due to changes in lighting, cross-modal fusion uses text as a supplement to enhance object pixel features, which can help segmentation prediction.

Method / Unseen Class	w/o Text		Full Model	
	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
bottle	77.3	51.6	**82.4**	**56.3**
coffee	74.7	64.3	**81.3**	**71**
lego	69.5	66.7	**84.2**	**80.7**
whisky	90.4	91.8	**95.2**	**96.4**
toaster	65	31.4	**68.4**	**35.6**
plush toy	69.2	65.8	**78**	**68.9**

Table 1. Ablation study to verify the role of text information.

As shown in Table 1, the performance of truly unseen classes dramatically increases.

*Equal contribution. This work was done when G. Li visited to Feng Zheng Lab in Southern University of Science and Technology.
†Corresponding author.

Figure 1. Visualization results on Mini-Ref-YouTube-VOS.

A man in a white shirt standing on the left is talking

A man in a brown jacket is talking

A man in black is talking

a man in a blue shirt

A man in a helmet is talking

A man in black standing on the right is talking

A man in purple is looking at his phone

A man in black is gesturing something with his hands

Figure 2. Visualization results on Mini-Ref-SAIL-VOS.