OnlineRefer: A Simple Online Baseline for Referring Video Object Segmentation
Supplementary Material

Dongming Wu[1‡], Tiancai Wang[2], Yuang Zhang[3], Xiangyu Zhang[2,4], Jianbing Shen[5†]
[1] Beijing Institute of Technology, [2] MEGVII Technology, [3] Shanghai Jiao Tong University,
[4] Beijing Academy of Artificial Intelligence, [5] SKL-IOTSC, CIS, University of Macau

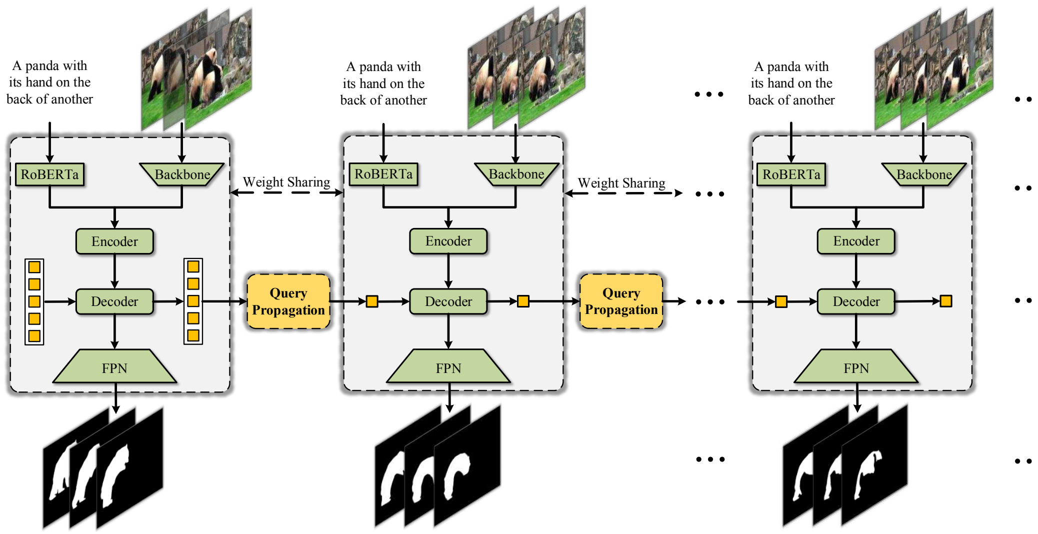

Figure 1: **Pipeline of semi-online model** of OnlineRefer, which associates the same referent object across different clips.

1. Semi-Online Model

We present more semi-online framework details of OnlineRefer in Fig. 1. Unlike the online model that follows a frame-by-frame pattern, the semi-online model propagates the target query across clips. Note that we use sharing query for multi-frame referring segmentation within each clip.

2. Additional Experiment Details

Refer-Youtube-VOS provides full-video expression by describing an entire video and first-frame expression based on the first frame, while we only use their full-video expression for training and validation.

Refer-DAVIS$_{17}$ also contains the full-video and first-frame expressions, which are developed by four annotators. Our final $\mathcal{J}\&\mathcal{F}$ scores are averaged from the four results.

3. Additional Ablation Study

We provide a thorough analysis of sampling length settings in Table 1. It is obvious that increasing frames from 2 to 3 brings performance improvement on ResNet-50, while it fails on Swin-L. When using the progressive sampling strategy (*i.e.*, [2, 3]), the best results can be obtained on both two backbones. This indicates that the appropriate increment on sampling lengths is beneficial for guaranteeing model stability and improving model performance.

Sampling Lengths	ResNet-50			Swin-L		
	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$	$\mathcal{J}\&\mathcal{F}$	$\mathcal{J}$	$\mathcal{F}$
[2]	56.5	55.1	57.9	62.7	60.8	64.5
[3]	56.7	55.2	58.2	/	/	/
[2, 3]	**57.3**	**55.8**	**58.8**	**63.5**	**61.6**	**65.5**
[2, 3, 4]	57.0	55.5	58.4	63.1	61.1	65.1

Table 1: **The effect of sampling lengths** on Refer-Youtube-VOS. '/' means no results due to model divergence.

4. More Qualitative Results

Fig. 2 offers some qualitative comparison between the offline method ReferFormer [1] and our OnlineRefer. We can see that OnlineRefer performs better than ReferFormer under the situations of object occlusion and visually-similar background, approving the superiority of query propagation. Fig. 3 also shows OnlineRefer can deal well with other challenging situations, like size and appearance variation, small or missing objects, moving objects, *etc*.

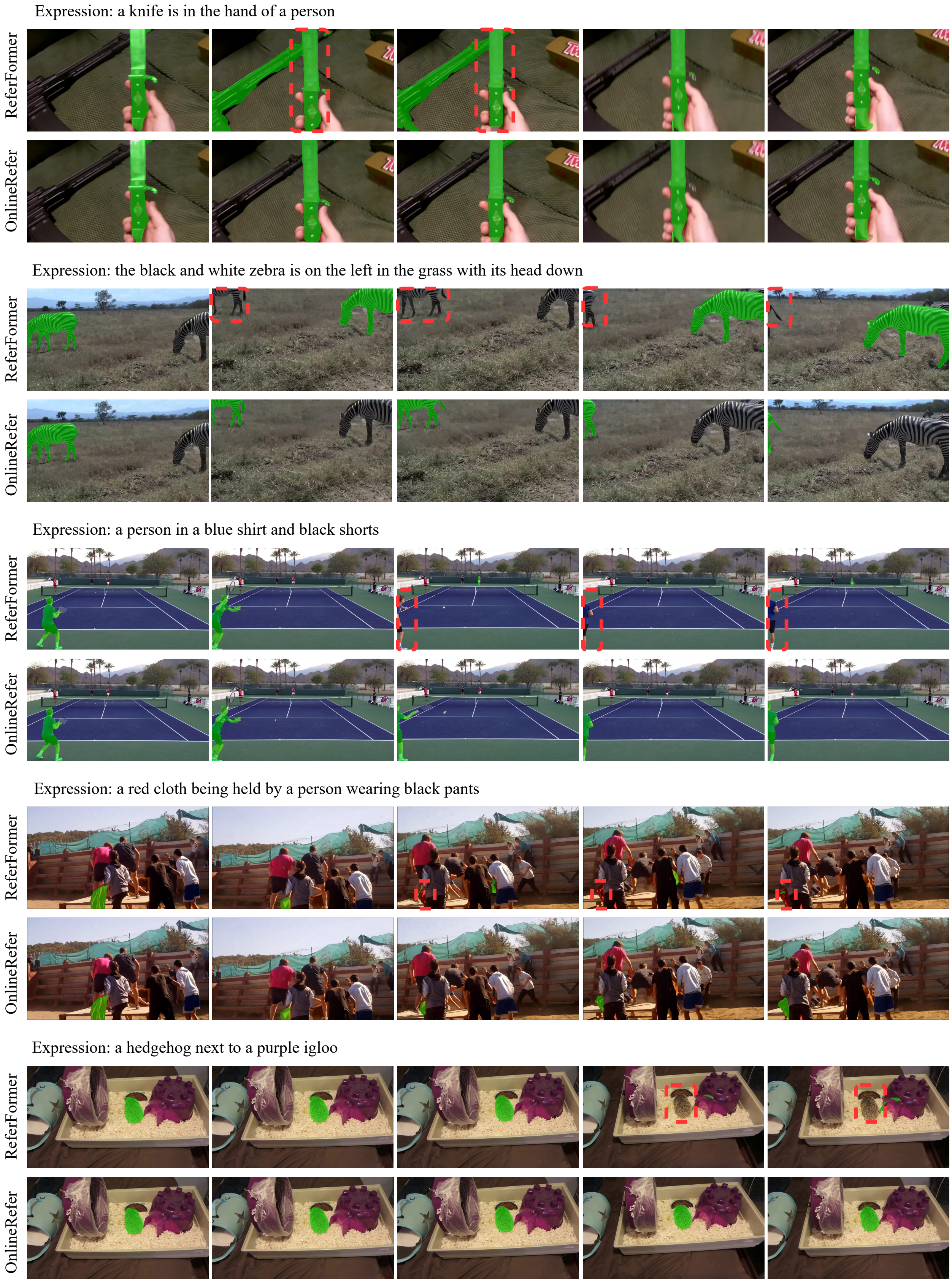

Figure 2: **Comparison between ReferFormer and OnlineRefer** on Refer-Youtube-VOS.

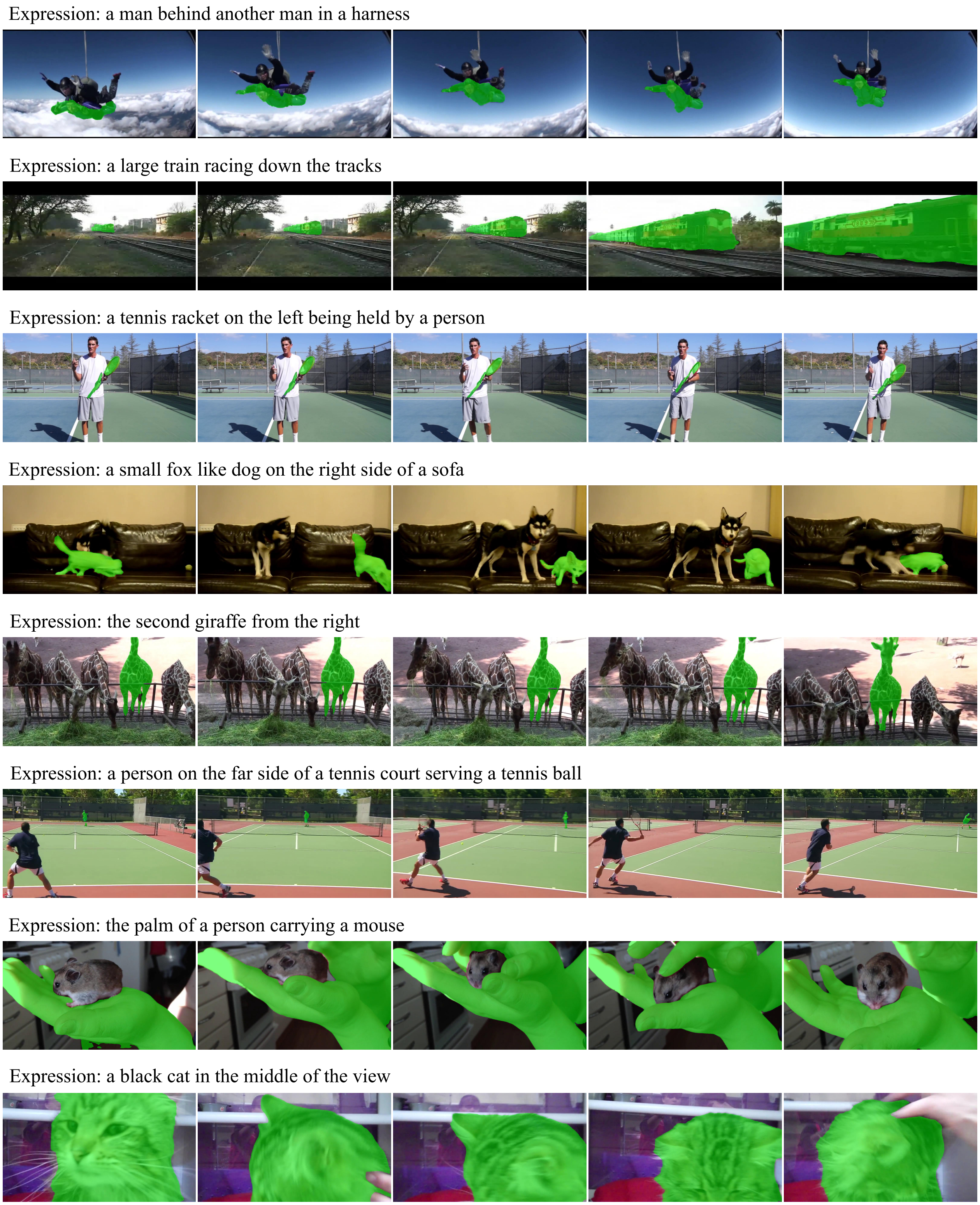

Figure 3: **More qualitative results** of our OnlineRefer on Refer-Youtube-VOS.

References

[1] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *CVPR*, 2022.