HTML: Hybrid Temporal-scale Multimodal Learning Framework for Referring Video Object Segmentation - Supplementary Material

Mingfei Han[1,4], Yali Wang[†2,6], Zhihui Li[3], Lina Yao[4], Xiaojun Chang[1,5], Yu Qiao[6,2]
[1]ReLER, AAII, UTS [2]Guangdong Provincial Key Laboratory of Computer Vision and Virtual Reality Technology, SIAT-SenseTime Joint Lab, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences [3]Shandong Artificial Intelligence, Qilu University of Technology
[4]Data61, CSIRO [5]Department of Computer Vision, Mohamed bin Zayed University of Artificial Intelligence [6]Shanghai AI Laboratory, Shanghai, China
https://mingfei.info/HTML/

1. More Implemented Details

For given videos during training, we use sliding windows to sample input video clips in a predefined frame number, *i.e*., the total frame number in our paper. All frames are downsampled by the short side to 360 and meanwhile ensure the maximum size of the long side is 640. Then, we apply the design Hybrid Temporal Scale Construction module to build hybrid temporal scales in a sequential manner, as in Sec.3.2.1 in the main submission. Take 8 total input frames as an example, the result temporal scales have 8, 5 and 3 frames in temporal scale 0, 1 and 2 respectively. Subsequently, the multimodal learning paths are constructed for each temporal scale simultaneously, with the parameters shared. During testing, we use a single temporal scale, which incurs no additional parameters or computational costs.

We train the model using AdamW optimizer [2] with a learning rate set to 5×10^{-5} for visual backbone and 10^{-4} for the rest and weight decay set to 5^{-4}. For a fair comparison with the previous state-of-the-art methods, we utilize the models pretrained with Ref-COCO [7], Ref-COCOg [7] and Ref-COCO+ [3] with total input number set to 1. We run the pretrain procedure for 12 epochs with the learning rate decaying by 10 at epochs 8 and 10, as in [6]. The box loss we utilized sums up the L1 loss and GIoU loss [5] with coefficients set to 5 and 2. The mask loss we utilized sums up the binary mask focal loss and DICE loss [4] with coefficients set to 1 and 1. The classification loss we utilized is the focal loss [1] with a coefficient set to 2.

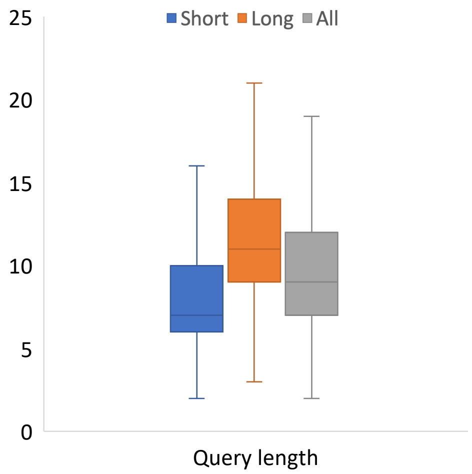


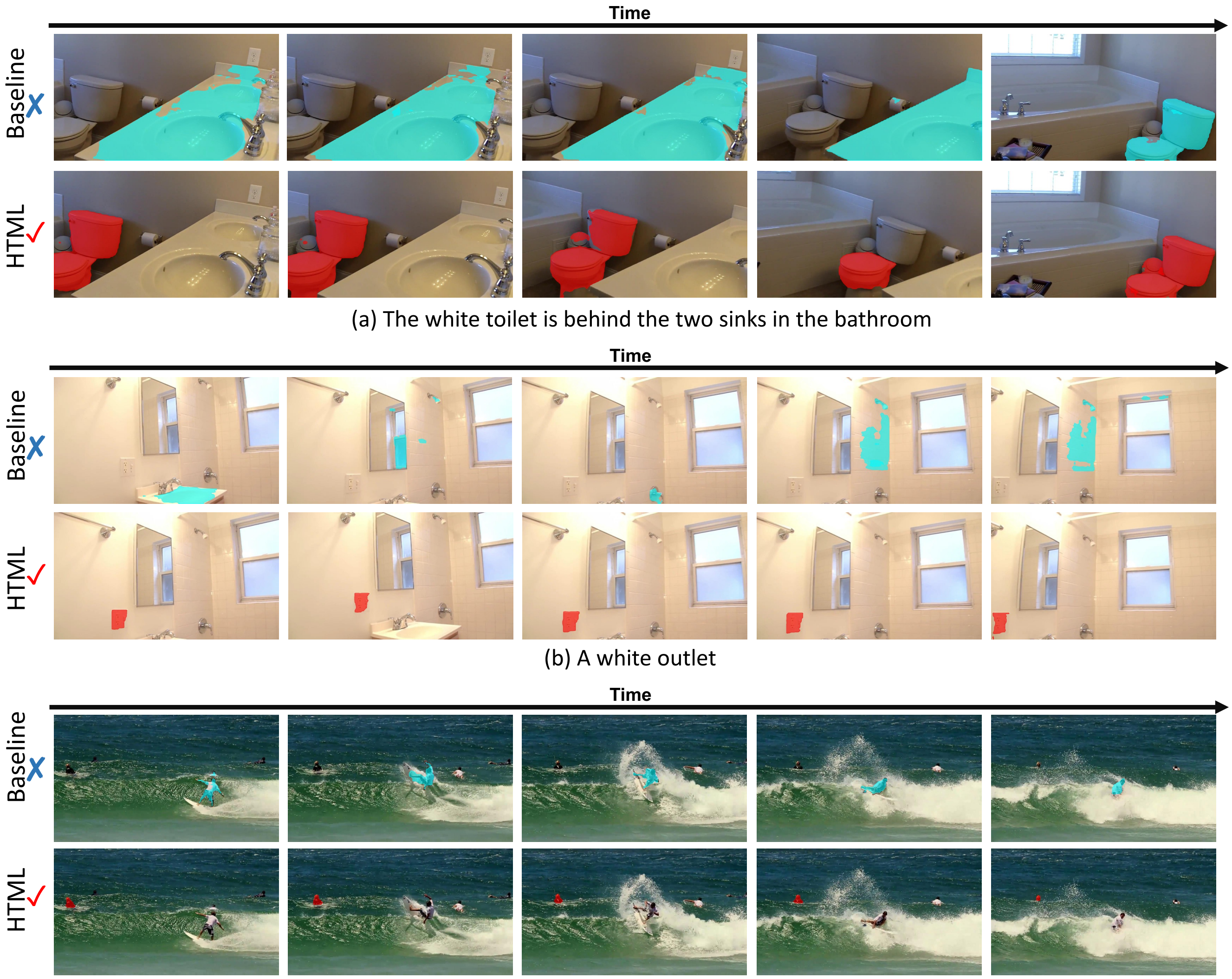
Figure 1: Statistics of different query sets, complementing Fig.3 in the main submission. The language descriptions to same object can be either short or long, containing simple or complex object semantics. In order to show our improved ability of handling diversified text expressions, we sample the shorter descriptions of the object to have the Short set and the longer descriptions to have the Long set for experiments .

2. Experiments

2.1. Length of language descriptions

As in Fig. 1, we show the statics of different query sets utilized in Sec.4.4 for ablation on the length of language descriptions. The median value of the number of words in the Short set is 7, fewer than that of the Long set and All set (11 and 9 respectively). As for the third quartile, the same regular is observed, with values of the three sets as

† Corresponding author.



(a) The white toilet is behind the two sinks in the bathroom

(b) A white outlet

(c) A person floating in the water to the left of a surfer wearing a white shirt

Figure 2: Visualization results of complex and simple language descriptions on Ref-Youtube-VOS. Red masks indicate positive segmentation results and blue masks indicate negative. Our proposed HTML can clarify the object confusion in both cases when similar objects, confusing backgrounds and complex scene structures exist.

10, 14 and 12 respectively. It proves that the text expressions in the Short set are shorter than that of the Long set, which is intuitively less complex. Further, it can be straightforwardly inferred that combining Short and Long sets, the All set posses much more diversity in language semantics. Finally, it can be proved that the three sets we used are adequate for validating the effectiveness of our method in exploring the flexibility and diversity in language descriptions for RVOS.

2.2. Visualization

We visualize the results of complex and simple language descriptions of Ref-Youtube-VOS in Fig. 2. Specifically, we compare two settings, *i.e.*, baseline with only simple temporal scale, and our HTML with hybrid temporal-scale learning capacity. As expected, our HTML succeeds in all three cases and when only a single temporal scale is applied during training, the model fails to segment the target object in different behaviours. In Fig. 2 (a), the toilet is referred to with the assistance of sinks. When only a single temporal scale is applied, the model can only successfully segment the object under the circumstance that the sinks are absent. Our HTML can succeed in all the frames by fully utilizing the clue of relative position to the sinks. In Fig. 2 (b), the outlet shares similar appearance with the background, *i.e.*, all white and with few textures. Single temporal scale fails to segment the object and regards the other white pixel cluster as the target. Differently, our HTML can fully discover the core object semantics by the interaction of text expression and different temporal scales and observing the tiny visual dynamics. In Fig. 2 (c), the *floating man* is in tiny scale and referred with the assistance of *surfer*, like the case shown in (a). The baseline model trained by a single temporal scale only fails in all frames, misled by the surfing man. Our HTML can discover the object semantics in the language description by engaging different temporal scales into the multimodal learning process and can successfully segment the object in all frames. All the results indicate that our HTML can effectively alleviate object confusion by dynamically constructing multimodal relations across temporal scales.

References

[1] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, pages 2980–2988, 2017. 1

[2] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[3] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 1

[4] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016. 1

[5] Hamid Rezatofighi, Nathan Tsoi, JunYoung Gwak, Amir Sadeghian, Ian Reid, and Silvio Savarese. Generalized intersection over union: A metric and a loss for bounding box regression. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 658–666, 2019. 1

[6] Jiannan Wu, Yi Jiang, Peize Sun, Zehuan Yuan, and Ping Luo. Language as queries for referring video object segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4974–4984, 2022. 1

[7] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *European Conference on Computer Vision*, pages 69–85. Springer, 2016. 1