Supplementary Material for Collaborative Spatial-Temporal Modeling for Language-Queried Video Actor Segmentation

Tianrui Hui[1,2*] Shaofei Huang[1,2,5*] Si Liu[3†] Zihan Ding[3] Guanbin Li[4,7]
Wenguan Wang[6] Jizhong Han[1,2] Fei Wang[5]
[1] Institute of Information Engineering, Chinese Academy of Sciences
[2] School of Cyber Security, University of Chinese Academy of Sciences
[3] Institute of Artificial Intelligence, Beihang University
[4] School of Computer Science and Engineering, Sun Yat-sen University
[5] SenseTime Research [6] Computer Vision Lab, ETH Zurich [7] Pazhou Lab, Guangzhou

1. Datasets

We conduct experiments on two popular language-queried video actor segmentation benchmark datasets. Details of the two datasets are summarized as follows.

A2D Sentences [2] is an extension of the Actor-Action Dataset [5] (A2D) with natural language descriptions for each video. It contains 8 actions categories performed by 7 actors categories with a total number of $3,782$ videos collected from YouTube. Each video has 3 to 5 frames of dense pixel-level annotations of actors and actions for training and evaluating segmentation performance. There are $6,655$ sentences in total which describe the actors and actions contained in each video. We follow the split of [2] to use $3,017$ training videos, 737 testing videos and 28 unlabeled videos.

J-HMDB Sentences [2] is extended from the J-HMDB dataset [3] which contains 21 different actions, 928 videos and corresponding 928 sentences. All the actors in J-HMDB dataset are humans which are annotated with 2D articulated human puppet masks for segmentation. For each video, one natural language query is annotated to describe the actions performed by the actors.

2. Network Training

After obtaining the outputs of our spatio-temporal decoder, we append a mask head containing two convolutions on each V_D^i to generate foreground response map P^i of the actor. In particular, we upsample V_D^1 to the original size of input frame to form P^0. Given the ground-truth binary mask Y^i at the i-th stage ($i \in [0, 5]$), the segmentation loss of our model is formulated as:

$$\mathcal{L}^i = \frac{1}{H^i W^i} \sum_{j=1}^{H^i} \sum_{k=1}^{W^i} \ell(P^{ijk}, Y^{ijk}), \tag{1}$$

where ℓ is the weighted binary cross entropy defined as:

$$\begin{aligned} \ell(P^{ijk}, Y^{ijk}) = &- \alpha Y^{ijk} log(\sigma(P^{ijk})) \\ &- (1 - Y^{ijk}) log(1 - \sigma(P^{ijk})). \end{aligned} \tag{2}$$

In the above equation, σ denotes $Sigmoid$ function and α is the weight of foreground pixels. Ground-truth masks of different resolutions are generated with nearest interpolation from original mask Y^0. The final loss of our model is the combination of 6 losses as follows:

$$\mathcal{L} = \sum_{i=0}^{5} \beta^i \mathcal{L}^i, \tag{3}$$

where β^i is the coefficient for the i-th loss. In our implementation, β^0 for loss $\mathcal{L}^0$ is 1.0 and others are set as 0.1. The weight of the foreground pixels α is 1.5.

3. Quantitative Results

3.1. Generalization results on J-HMDB

We compare our method with ACGA [4] (only its code is released) over JHMDB under three settings. (i) *full supervision*: Training and testing on J-HMDB Sentences. (ii) *w/ fine-tuning*: First training on A2D Sentences, then fine-tuning and testing on J-HMDB Sentences. (iii) *w/o fine-tuning*: First training on A2D Sentences, then testing on J-HMDB Sentences. In term of Mean IoU (mIoU), our method (68.5%, 69.8% and 59.7%) significantly outperforms ACGA (63.4%, 66.7% and 58.5%), demonstrating our method has better generalization ability.

*Equal contribution
†Corresponding author

3.2. Non-trivial subset of A2D Sentences

We also evaluate our model on the *non-trivial* subset of A2D (re-categorized by RefVOS [1]), where *non-trivial* means the referent (both the object and its referring expression) is not the only object of its class in the video. Our model achieves 42.3% mIoU on *non-trivial*, significantly outperforming RefVOS by 9.1%. This clearly shows our model can distinguish the referent between multiple actors of the same class.

3.3. Temporal Consistency

We devise a consistency metric *Video Precision@X* (VP@X), where a testing video is considered as correct if all the annotated frames in this video have IoU scores higher than the threshold X. The ratio between the numbers of correct testing videos and total testing videos is VP@X. Our model achieves 56.0% VP@0.5 while ACGA achieves 49.3%, indicating our model can predict steady masks along the entire video.

3.4. Ablations on LGFS and CMAM

Inserting LGFS into $\{\}$, $\{I_5\}$ and $\{I_5, I_4, I_3\}$ yield 55.3%, 55.8% and 56.0% mIoU respectively, showing LGFS works better in deeper stages. Similar to the settings in Table 3(a) of our main text, *Spa+CMAM, Temp+CMAM and Spa+Temp+CMAM* yield 53.8%, 52.6% and 55.3% mIoU respectively, showing CMAM is more effective in our model. But LGFS also has its own gains over the strong baseline with CMAM, forming a helpful component.

4. Qualitative Results

4.1. Visualization of LGFS

As shown in Figure 1, for the left case, temporal channel produces higher responses for the crawling child and obtains larger selection weight than spatial channel (0.9904 *vs.* 0.0096). For the right case, spatial channel becomes dominant for the static bird on left (the right bird just flied onto the wall). These visualization results illustrate the adaptive selection ability of LGFS.

4.2. Comparison with other methods

In Figure 2, we present the qualitative comparison between ACGA [4] and our method on the A2D Sentences dataset. Different colors of the queries correspond to different segmentation masks in each frame. From Figure 2 we can observe that our method can yield more complete and coherent segmentation masks on the queried actors while ACGA only generates partial prediction on the actors. For example, the person in row (a) and the ball in row (c) are segmented with incomplete and oversized masks. In addition, ACGA also tends to misidentify the queried actor.

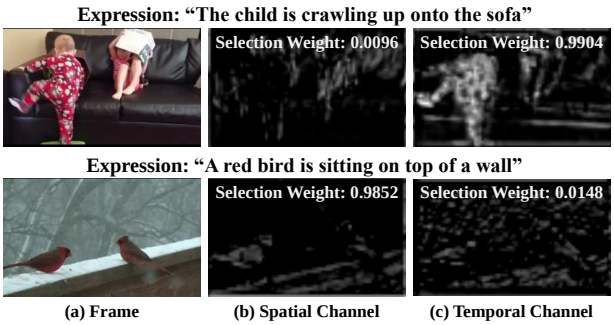

Figure 1. Attention maps of LGFS.

As shown in row (e), the dog is misidentified as the toddler, which demonstrates the effectiveness of our method on spatial-temporal multimodal modeling.

References

[1] Miriam Bellver, Carles Ventura, Carina Silberer, Ioannis Kazakos, Jordi Torres, and Xavier Giro-i Nieto. Refvos: A closer look at referring expressions for video object segmentation. *arXiv preprint arXiv:2010.00263*, 2020. 2

[2] Kirill Gavrilyuk, Amir Ghodrati, Zhenyang Li, and Cees GM Snoek. Actor and action video segmentation from a sentence. In *CVPR*, 2018. 1

[3] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J Black. Towards understanding action recognition. In *ICCV*, 2013. 1

[4] Hao Wang, Cheng Deng, Junchi Yan, and Dacheng Tao. Asymmetric cross-guided attention network for actor and action video segmentation from natural language query. In *ICCV*, 2019. 1, 2, 3

[5] Chenliang Xu, Shao-Hang Hsieh, Caiming Xiong, and Jason J Corso. Can humans fly? action understanding with multiple classes of actors. In *CVPR*, 2015. 1

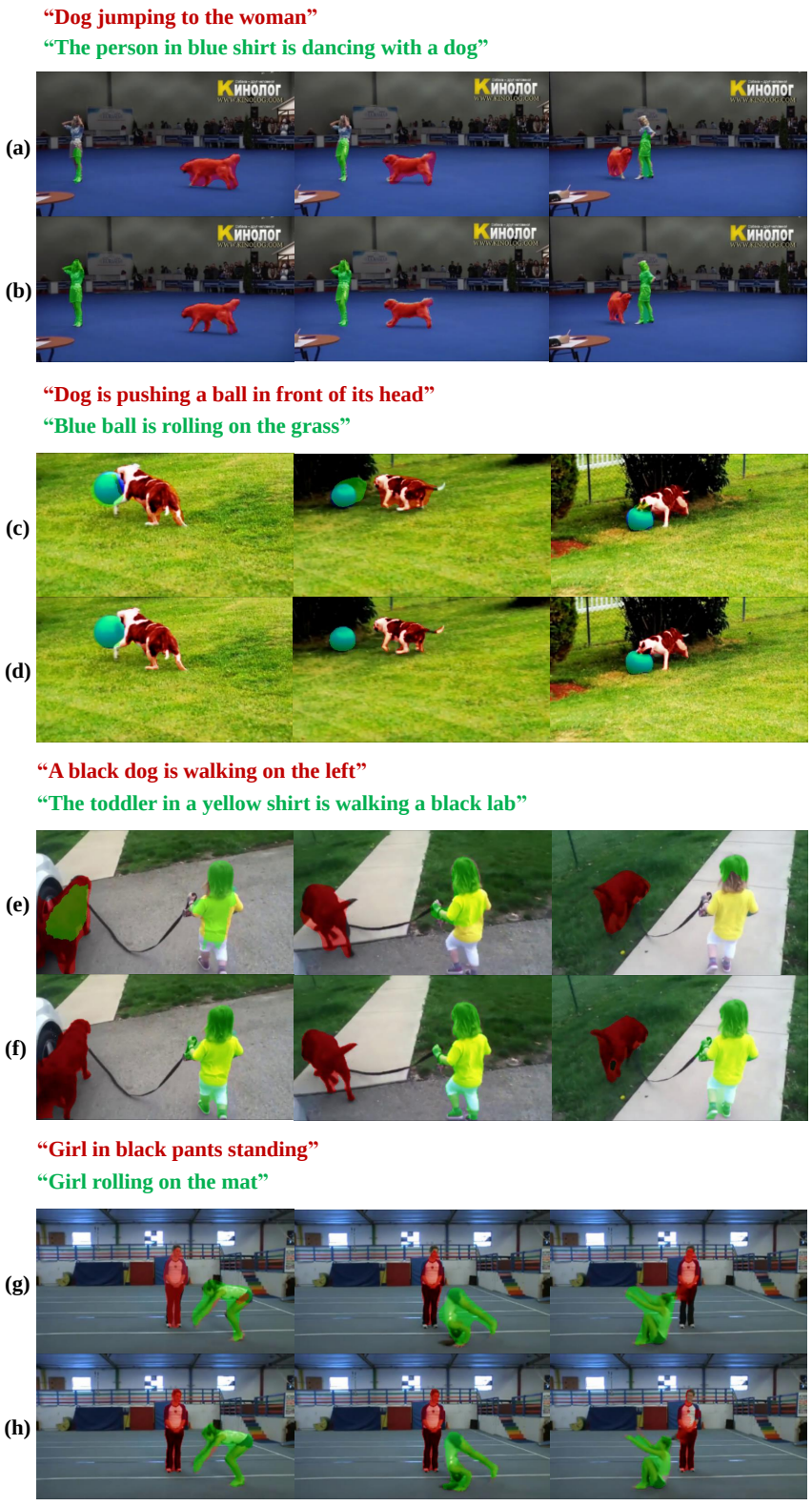

Figure 2. Qualitative results on A2D Sentences. (a)(c)(e)(g) Results of ACGA [4]. (b)(d)(f)(h) Results of our model.