Semantic and Sequential Alignment for Referring Video Object Segmentation

Supplementary Material

In this supplementary material, we present additional experiments, as well as more details of the proposed framework SSA and visualization results, as follows:

- In Sec. A, we firstly supplement experiments on A2D-Sentences and JHMDB-Sentences datasets [10] to evaluate the performance of SSA.
- We further introduce the implementation details of SSA including model structure and inference pipeline, which are demonstrated in Sec. B.
- Finally, we provide more visualization results on the MeViS dataset [7] in Sec. C.

A. Additional Experiments

Here, to further explore the generalization ability of our framework SSA, we conduct additional experiments on the other two RVOS datasets: A2D-Sentences and JHMDB-Sentences [10].

A.1. More Datasets Evaluations

Datasets. A2D-Sentences and JHMDB-Sentences are created by providing the additional textual annotations on the original A2D and JHMDB datasets [10]. The A2D-Sentences dataset contains 3,782 videos and each video has 3-5 annotated segmentation masks and JHMDB-Sentences totally comprises 928 videos, each of which is associated with a text description.

Evaluation Metrics. Following previous work [54], the model is evaluated with the criteria of Precision@K, Ovrall IoU, Mean IoU and mAP over 0.50:0.05:0.95. The Precision@K measures the percentage of test samples whole IoU scores are higher than the threshold K. Following standard protocol, the thresholds are set as 0.5:0.1:0.9.

A.2. Experimental Settings

We adopt a similar training strategy following DsHmp [13]. For A2D-Sentences, We first conduct pre-training on the image-level datasets RefCOCO/+/g [34, 58], which lasts for 100,000 iterations. Then we train our model on A2D-Sentences for 30,000 iterations. For JHMDB-Sentences, we directly apply the learned model from A2D-Sentences to JHMDB-Sentences without finetuning.

Notably, video samples in A2D-Sentences dataset do not include annotations for all frames. Therefore,we only select the single annotated frame in each video for training and inference, meaning that the number of processed frame during both training and inference is 1. The same strategy is applied to the evaluation on JHMDB-sentences dataset.

A.3. Results

Our method still achieves state-of-the-art performance on both the A2D-Sentences and JHMDB-Sentences datasets, even under the single-frame training and inference setting, which demonstrates the strong generalization capability of SSA.

A2D-Sentences. As shown in Tab. 8, on A2D-Sentences dataset, we achieve competitive results with **58.5**% mAP, **80.7**% Overall IoU, and **72.9**% Mean IoU, outperforming the best method LoSh [60], by **0.9**%, **1.4**%, and **1.3**%, respectively. And SSA also demonstrates its strong ability to produce high-quality masks via the stringent metrics (*e.g.* 65.6% for P@0.8 and 29.3% for P@0.9, improving of 7.9% and 7.5% over SOC [33]).

JHMDB-Sentences. As shown in Tab. 9, on JHMDB-Sentences dataset, we also achieve new state-of-the-art results with **45.9**% mAP, **73.7**% Overall IoU, and **72.5**% Mean IoU, which surpass the current method DsHmp [13] by **1.0**%, **0.6**%, and **0.4**%, respectively. This highlights the strong generalization capability of SSA.

B. Implementation Details

B.1. Model Structure

The CLIP image and text encoders used in Sec. 3.2 are pre-trained on LAION-2B [46] from OpenCLIP. Instance query generation Module (Sec. 3.3) consists of encoder with 6 layers and decoder with 9 layers, aligning with the standard Mask2Former structure [5]. Specifically, we implement bidirectional cross-attention interactions (Eq. (5)) within each layer of the encoder to generate instance queries. Video Decoder (Sec. 3.4) employs six layers, each incorporating cross-attention layer, self-attention layer and FFN. The number of frame queries Q is set to $N = 20$. Consequently, video queries Q_v initialized from them is also set to 20.

B.2. Inference Pipeline

In Sec. 3.4, we obtain the processed video queries Q_{emb} and logits S_{cls} from video decoder. When addressing reference involving multiple targets (*e.g.* MeViS [7]), we select the Q_{emb} with S_{cls} greater than a specified threshold σ as $\hat{Q}_{emb}$. While for single-target reference (*e.g.* Ref-Youtube-VOS [47] and A2D-Sentences [10]), we directly obtain the $\hat{Q}_{emb}$ with the highest S_{cls} through argmax function:

$$\hat{Q}_{emb} = \begin{cases} \{Q^i_{emb} \mid S^i_{cls} > \sigma\}, \textit{if multi-target} \\ Q^{\arg\max(S_{cls})}_{emb}, \textit{otherwise} \end{cases} \in \mathbb{R}^{N' \times C}, \tag{9}$$

Table 8. **Comparison with state-of-the-art models on A2D-Sentences dataset [10].**

Methods	Backbone	Precision					IoU		mAP
		P@0.5	P@0.6	P@0.7	P@0.8	P@0.9	Overall	Mean	
LBDT [8]	ResNet-50	73.0	67.4	59.0	42.1	13.2	70.4	62.1	47.2
MTTR [4]	Video-Swin-T	75.4	71.2	63.8	48.5	16.9	72.0	64.0	46.1
ReferFormer [54]	Video-Swin-T	82.8	79.2	72.3	55.3	19.3	77.6	69.6	52.8
OnlineRefer [53]	Video-Swin-B	83.1	80.2	73.4	56.8	21.7	79.6	70.5	-
HTML [12]	Video-Swin-T	82.2	79.2	72.3	55.3	20.1	77.6	69.2	53.4
SgMg [35]	Video-Swin-T	-	-	-	-	-	78.0	70.4	56.1
TempCD [49]	ResNet-50	-	-	-	-	-	76.6	68.6	-
SOC [33]	Video-Swin-T	83.1	80.6	73.9	57.7	21.8	78.3	70.6	54.8
LoSh [60]	Video-Swin-T	-	-	-	-	-	79.3	71.6	57.6
DsHmp [13]	Video-Swin-T	-	-	-	-	-	79.0	71.3	57.2
Ours	CLIP	**84.8**	**83.2**	**78.4**	**65.6**	**29.3**	**80.7**	**72.9**	**58.5**

Table 9. **Comparison with state-of-the-art models on JHMDB-Sentences dataset [10].**

Methods	Backbone	Precision					IoU		mAP
		P@0.5	P@0.6	P@0.7	P@0.9	P@0.9	Overall	Mean	
LBDT [8]	ResNet-50	86.4	74.4	53.3	13.2	0.0	64.5	65.8	41.1
MTTR [4]	Video-Swin-T	93.9	85.2	61.6	16.6	0.1	70.1	69.8	39.2
ReferFormer [54]	Video-Swin-T	95.8	89.3	66.8	18.9	0.2	71.9	71.0	42.2
OnlineRefer [53]	Video-Swin-B	96.1	90.4	71.0	21.9	0.2	73.5	71.9	-
HTML [12]	Video-Swin-T	-	-	-	-	-	-	-	42.7
SgMg [35]	Video-Swin-T	-	-	-	-	-	72.8	71.7	44.4
TempCD [49]	ResNet-50	-	-	-	-	-	70.6	69.6	-
SOC [33]	Video-Swin-T	96.3	88.7	67.2	19.6	0.1	72.7	71.6	42.7
DsHmp [13]	Video-Swin-T	-	-	-	-	-	73.1	72.1	44.9
Ours	CLIP	**96.9**	**92.5**	**73.2**	**22.4**	0.1	**73.7**	**72.5**	**45.9**

where N' denotes the number of filtered video queries.

The filtered mask embeddings $\hat{Q}_{emb}$ are then multiplied with the mask features $\mathcal{F}_{mask}$ to obtain the final predicted masks:

$$\mathcal{M} = \text{sigmoid}(\mathcal{F}_{mask} \cdot \hat{Q}_{emb}), \tag{10}$$

where $\mathcal{M} \in \mathbb{R}^{T \times H \times W}$, denotes the binary masks of the referred target(s).

C. More Visualization Results

In this section, we provide additional visualization results on the MeViS dataset [7], which involves extensive motion descriptions, making it more reflective of real-world scenarios. We find that our SSA framework demonstrates superior perception capability for scenarios frequently encountered in the real world, such as non-prominent targets, complex motion environments, and multi-object descriptions.

DsHmp [13] still struggles to perfectly handle the above situations. For instance, the first row in Fig. 7(a) demonstrates that DsHmp only understands "gripping the pole" but failed to capture the subsequent action of "body facing downward", leading to incorrect target segmentation. And in Fig. 7(b), when faced with complex environment, such as the grazing sheep hidden in a dark corner with many similar objects in front, DsHmp is distracted by other sheep, leading to an additional erroneous segmentation (1^{st} row). Finally, in Fig. 7, DsHmp failed to track the two girls appeared in the middle frames of the video accurately, missing one of them (1^{st} row). We attribute this issue to the insufficient modeling of semantic and sequential consistency.

In contrast, SSA leverages the powerful features from semantic alignment and the instance modeling from sequential alignment, enabling precise multi-modality understanding. As shown in Fig. 7, SSA can accurately perceive comprehensive action descriptions (the 2^{nd} row in Fig. 7(a)), non-prominent targets in complex environments (the 2^{nd} row in Fig. 7(b)), and scenarios involving the emergence and disappearance of multiple targets (the 2^{nd} row in Fig. 7(c)).

These visualization results further emphasize the importance of **semantic alignment** and **sequential alignment** for real-world referring video object segmentation.

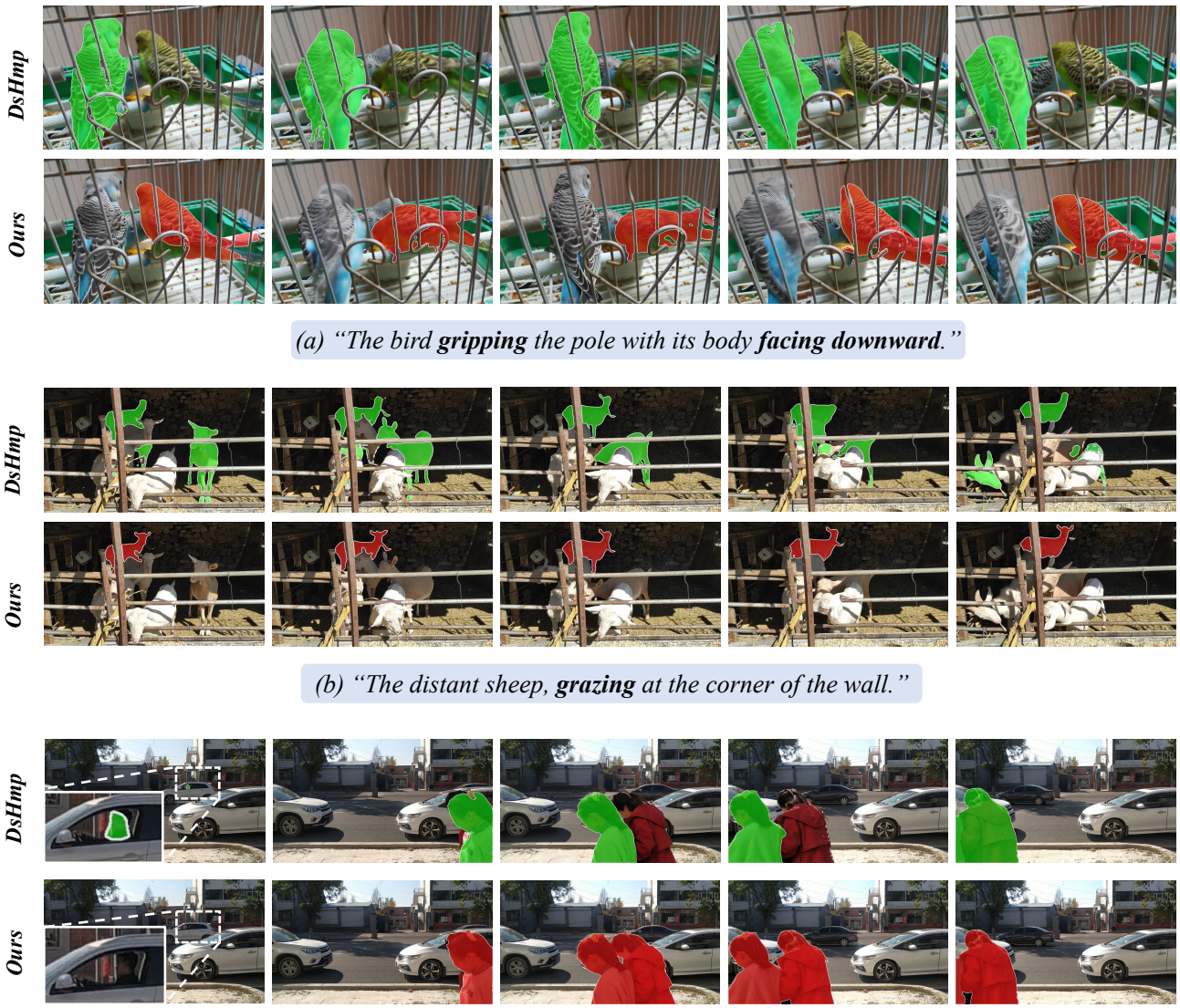

Figure 7. More visualization results on Mevis dataset [7]. The first row indicates the segmentation results of DsHmp [13] while the second row indicates the segmentation results of ours.