LoSh: Long-Short Text Joint Prediction Network for Referring Video Object Segmentation

Supplementary Material

Method	Backbone	O-IoU	M-IoU	mAP
Hu *et al*. [3]	VGG-16	54.6	52.8	17.8
Gavrilyuk *et al*. [2]	I3D	54.1	54.2	23.3
CMPC-V [5]	I3D	61.6	61.7	34.2
ClawCraneNet [4]	ResNet-50/101	64.4	65.6	-
ReferFormer [7]	Video-Swin-T	71.9	71.0	42.2
ReferFormer [7]	Video-Swin-B	73.0	71.8	43.7
MTTR ($w = 8$) [1]	Video-Swin-T	67.4	67.9	36.6
LoSh-M ($w = 8$)	Video-Swin-T	**70.8**	**70.4**	**39.0**
SgMg($w = 5$) [6]	Video-Swin-T	72.8	71.7	44.4
LoSh-S($w = 5$)	Video-Swin-T	**73.6**	**72.5**	**45.0**
SgMg($w = 5$) [6]	Video-Swin-B	73.7	72.5	45.0
LoSh-S($w = 5$)	Video-Swin-B	**74.5**	**73.4**	**45.7**

Table 1. Quantitative comparison with state of the art on JHMDB-Sentences [2]. O-IoU and M-IoU represent Overall IoU and Mean IoU. The number of input frames w follows the implementation details of [1, 6].

In this supplementary material, we provide 1) quantitative comparisons between LoSh and other methods on the JHMDB-Sentences dataset; 2) additional ablation studies for LoSh-M and LoSh-S; 3) additional qualitative results of mask predictions across frames. The experimental setup follows the default setting in the main body of our paper.

A. Additional quantitative comparisons with state of the art on the JHMDB-Sentences

To further show the generalizability of our method, we follow [1, 6, 7] to evaluate the trained LoSh-M and LoSh-S from the A2D-Sentences onto the JHMDB-Sentences without fine-tuning. As shown in Tab. 1, LoSh-M and LoSh-S gain massive improvements on all metrics compared to their counterpart baselines (*e.g.*, +2.4 mAP, +3.4% Overall IoU and +2.5% Mean IoU when comparing between LoSh-M and MTTR). Furthermore, LoSh-S with Video-Swin-B yields highest results amongst all.

B. Additional ablation studies on LoSh-M

For a fair comparison, we train baseline MTTR [1] using original long expressions and our generated short ones with their corresponding GTs (Baseline w/ Sh in Tab. 2). It shows a relatively marginal increase of +0.6 mAP on A2D-Sentences. While our LoSh-M, by adding interactions between long and short expressions, achieves a noteworthy improvement of +3.1 mAP. We also provide more ablation studies on LoSh-M in terms of the number of input frames and object queries.

Number of input frames w. We study the effect of the number of input frames on LoSh-M in Tab. 3. Generally, a larger number of input frames helps the model better extract motion information across frames. Note that when $w = 1$, the forward-backward visual consistency loss is deprecated since we can not generate optical flow from only one input frame. When changing w from 1 to 8, we observe an mAP gain of 4.0 and a Mean IoU gain of 3.1%. Yet, when $w = 12$, the performance slightly drops; we suspect that when the input video gets long, the content becomes complex and irrelevant information is more likely to be included.

Method	IoU		mAP
	Overall	Mean	
Baseline	70.2	61.8	44.7
Baseline w/ Sh	72.3	63.5	45.4
Losh-M (Ours)	**72.9**	**64.9**	**47.8**

Table 2. Ablation study for long-short text joint prediction.

w	IoU		mAP
	Overall	Mean	
1	70.3	61.8	43.8
5	72.3	64.5	46.9
8	**72.9**	**64.9**	**47.8**
12	72.4	64.5	47.0

Table 3. Ablation study for the number of input frames.

N	IoU		mAP
	Overall	Mean	
20	72.6	64.2	45.8
50	**72.9**	**64.9**	**47.8**
80	72.4	64.2	45.5

Table 4. Ablation study for the number of object queries.

Number of object queries N. We study the effect of the number of object queries on LoSh-M in Tab. 4. Given our default setting $N = 50$, LoSh-M gains the best performance. The performance drops with a smaller N as the smaller set of object queries might not be diverse enough to cover the target instance in the video. However, a larger N also ends up with some performance drop. During training, only the object query which is matched with the ground-truth instance is trained while others are ignored. This means that many object queries cannot be fully trained, resulting into sub-optimal performance.

C. Additional ablation studies on LoSh-S

In this section, we provide basic ablation studies on LoSh-S with Video-Swin-T as visual encoder and RoBERTa as linguistic encoder in terms of our proposed components.

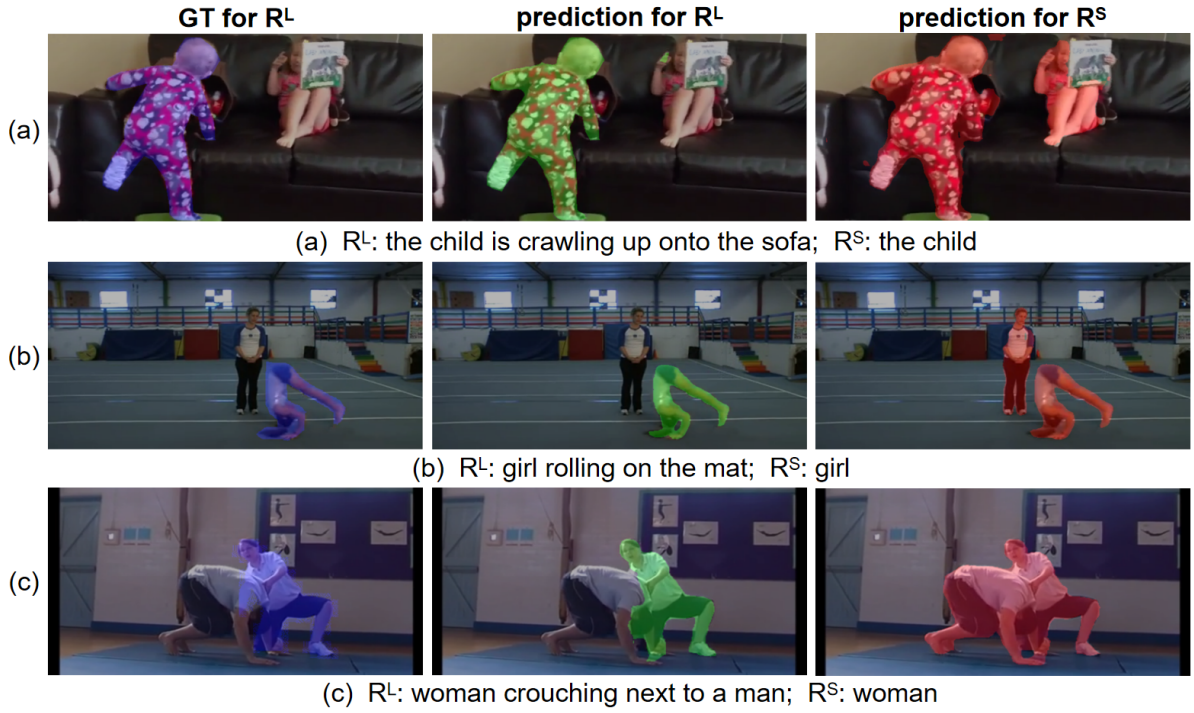

Figure 1. Qualitative results of mask predictions from LoSh using long and short text expressions, respectively. The three columns from left to right are ground truth, predictions for the long text expressions (R^L), and the short expressions (R^S), respectively. The long and short text expressions, R^L and R^S, are written below each row.

Figure 2. Qualitative results of LoSh across frames.

Method	IoU		mAP
	Overall	Mean	
LoSh-S w/o Sh	78.5	70.8	56.5
LoSh-S w/o $\mathcal{L}_{lsi}$	78.7	71.0	56.8
LoSh-S w/o CA	78.9	71.3	57.2
LoSh-S w/o $\mathcal{L}_{fbc}$	79.0	71.3	57.3
LoSh-S (Ours)	**79.3**	**71.6**	**57.6**

Table 5. Ablation study for the components in LoSh-S.

Long-short text joint prediction. Similar to Sec. 4.4 in the paper, we first present the result of LoSh-S using only the long text expressions without the short text expressions (LoSh-S w/o Sh), which is equivalent to the baseline SgMg [6] with the proposed forward-backward visual consistency. The result is reported in Tab. 5: compared to LoSh-S, LoSh-S w/o Sh has a clear performance drop on mAP and IoU, *e.g.*, -1.1 on mAP and -0.8% on both IoU. We then present the result of LoSh-S without the proposed long-short cross-attention modules, *i.e.*, LoSh-S w/o CA in Tab. 5. It shows a 0.4 decrease on mAP compared to LoSh-S. Last, we ablate the proposed long-short predictions intersection loss $\mathcal{L}_{lsi}$ by presenting a variant of LoSh-S without using $\mathcal{L}_{lsi}$, *i.e.*, LoSh-S w/o $\mathcal{L}_{lsi}$. The long-short text expressions and cross attention are still used. According to

Tab. 5, we observe a 1.3 decrease on mAP from LoSh-S to LoSh-S w/o $\mathcal{L}_{lsi}$. Without using $\mathcal{L}_{lsi}$, the model can not well align the predicted masks for the long and short text expressions.

Forward-backward visual consistency loss. We present the result of LoSh-S without using the forward-backward visual consistency loss, *i.e*. LoSh-S w/o $\mathcal{L}_{fbc}$ in Tab. 5. We observe a 0.3 decrease on mAP and 0.4% decreases on both Overall IoU and Mean IoU, compared to LoSh-S.

D. Additional qualitative results

Qualitative results for long-short text predictions. Compared with the long text expressions, the short ones are more generic expressions which normally contain only the appearance-related information of the subjects. Recalling to Sec. 4.2, there exists a few cases (approximately 10% in RVOS datasets) in which they refer to multiple instances in video clips. We visualize the respective mask predictions corresponding to long and short text expressions in these cases. According to Fig. 1, the mask prediction of LoSh using the short text expression tends to cover a broader potential area in the input video compared to that generated using the long text expression. Although the short text expressions refer to more instances in these examples, our LoSh can still generate reasonable mask predictions for them.

Qualitative results across frames. We show the qualitative results of LoSh across frames in Fig. 2. Our model can successfully segment the target instances corresponding to the input text expressions in challenging scenarios (*e.g*., partial disappearance in the camera, high-speed and frequent movement, variety of poses, occlusion).

References

[1] Adam Botach, Evgenii Zheltonozhskii, and Chaim Baskin. End-to-end referring video object segmentation with multimodal transformers. In *CVPR*, 2022. 1

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

[3] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In *ECCV*, 2016. 1

[4] Chen Liang, Yu Wu, Yawei Luo, and Yi Yang. Clawcranenet: Leveraging object-level relation for text-based video segmentation. *arXiv preprint arXiv:2103.10702*, 2021. 1

[5] Si Liu, Tianrui Hui, Shaofei Huang, Yunchao Wei, Bo Li, and Guanbin Li. Cross-modal progressive comprehension for referring segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4761–4775, 2021. 1

[6] Bo Miao, Mohammed Bennamoun, Yongsheng Gao, and Ajmal Mian. Spectrum-guided multi-granularity referring video object segmentation. In *ICCV*, 2023. 1, 2

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