Supplementary Materials for "SiRi: A Simple Selective Retraining Mechanism for Transformer-based Visual Grounding"

Mengxue Qu[1,2⋆], Yu Wu[3], Wu Liu[4], Qiqi Gong[1,2], Xiaodan Liang[5], Olga Russakovsky[3], Yao Zhao[1,2], and Yunchao Wei[1,2]

[1]Institute of Information Science, Beijing Jiaotong University
[2]Beijing Key Laboratory of Advanced Information Science and Network Technology
[3]Princeton University [4]JD Explore Academy [5]Sun Yat-sen University
qumengxue@bjtu.edu.cn, yuwu@princeton.edu, wychao1987@gmail.com

1 Details of Multi-task SiRi

As shown in Fig. 1, in multi-task SiRi, we leverage an auxiliary decoder (no weights sharing) for multi-task learning in each training/retraining stage. The losses of two decoders are summed up as the overall objective function for optimization.

After training/retraining, the auxiliary decoder was dropped after training so that we keep the same amount of parameters and operations (inference speed) in model inference.

In detail, we generate constant grid points by dividing the image into patches. Then we take the grid intersections for position encoding, as shown in Fig. 1. The coordinates of the k-th intersection point P_k are,

$$P_k = (\frac{k_1}{\sqrt{n}+1}, \frac{k_2}{\sqrt{n}+1}), k_1, k_2 \in \{1, 2, ..., \sqrt{n}\}, \tag{1}$$

where n is the number of object queries. Based on the generated constant points P, the constant queries Q_c can be formulated as follows,

$$Q_c = \begin{cases} PE(P, 2i) = sin(\frac{P}{10000^{2i/C}}) \\ PE(P, 2i+1) = cos(\frac{P}{10000^{2i/C}}), \end{cases} \tag{2}$$

where C denotes the dimension of the query embedding, and i is the dimension index.

Therefore, in multi-task SiRi, we leverage an auxiliary decoder (no weights sharing) for multi-task learning in each training/retraining stage. This auxiliary decoder was dropped after training so that we keep the same amount of parameters and operations in model inference.

During training, the model weights of the two decoders are randomly initialized and separately updated. In other words, they do not share weights. We

⋆ Work done during an internship at JD Explore Academy.

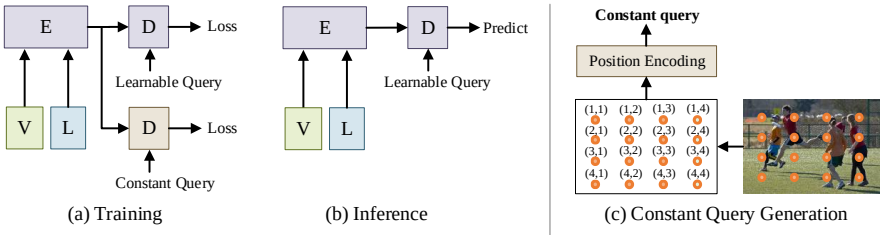


Fig. 1. Overview of the multi-task SiRi. We generate constant query as shown in (c).

individually calculate the loss on each decoder's prediction and then simply add the two losses as the overall objective function for optimization. For inference, we can keep *either one* of the two trained decoders and take its prediction as the final prediction. Thus, the inference speed is exactly the same as the previous single decoder framework.

We found in experiments that both decoders in the multi-task structure achieve very similar performance and significantly outperform either of them in the previous single-task framework. This proves that the performance gains are from better-optimized encoders, rather than additional computation or model parameters.

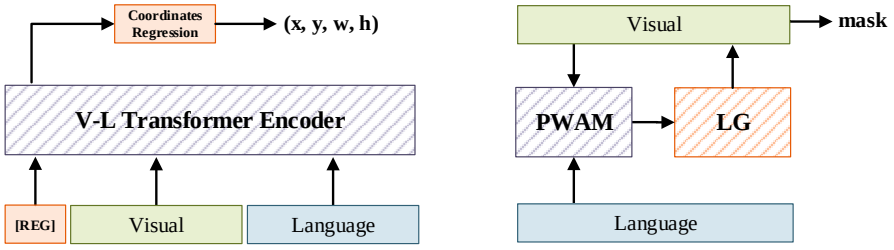
2 Additional Experimental Analysis

2.1 More Experiment Results and Setting Details of SiRi in other V-L models

As shown in Fig. 2 (a), we apply SiRi mechanism by keeping the V-L Transformer encoder continually trained while the other modules re-initialized. The [REG] token is used to enter the coordinates regression module and locate the object. We strictly adopted the same hyper-parameters as TransVG. In LAVT [2], Pixel-Word Attention Module (PWAM) and Language Gate (LG) are modules for V-L interaction. As shown in Fig. 2 (b), we keep the two module parameters continuously trained and reinitialize the other parameters.

2.2 Additional Qualitative Results

We show more qualitative results of our trained model in Fig. 4 and Fig. 5. Each example set includes the ground truth (the left one), prediction of our method (the middle one), and the attention map of the encoder (the right one). The green box indicates the ground truth annotation, while the red one represents the prediction box of our trained model. Fig. 4 shows some correct prediction examples of referring expression comprehension, while Fig. 5 contains several incorrect predictions. These visualization examples demonstrate that our approach can model the relative relationship description, *e.g.*, the relationship of



(a) SiRi in TransVG (b) SiRi in Lavt

Fig. 2. The illustration of (a) TransVG [1] and (b) LAVT [2] with SiRi. The solid color background means re-initializing, while the slash color background means continually updating.

"couch" and **"person"** in "couch under person in black". In addition, we can also find that the attention map of the encoder tends to be more attentive to the object referred to by the expression (with higher brightness).

2.3 Training Loss

For the error cases, we found the network usually fails if the referred object is obscured or occluded, *e.g.*, in "bread closer to bowl", the target object is occluded. Another common error case is that the referring expression is based on the text content on the object, *e.g.*, "happy birthday cake".

Fig. 3 depicts the loss curve of the training process using SiRi mechanism. The retraining period is set as 30 epochs. As can be seen from the loss curves, the model reaches a better local minimum after each retraining progress. It verifies our motivation that a better initialized encoder for vision-language perception usually helps the model converge to a better local minimum.

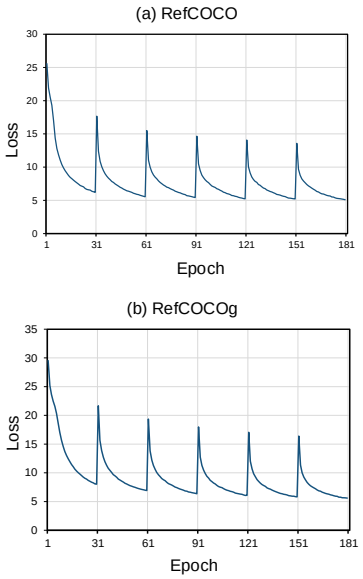


Fig. 3. Training loss of SiRi mechanism on (a) RefCOCO and (b) RefCOCOg using constant learning rate.

2.4 Comparison with Large Pre-training

We report SiRi *with* large pre-training in the table below. We can see that

SiRi could further improve even when large-scale pre-training has provided superior initialization for the whole model.

Model	RefCOCO			RefCOCO+		
	val	testA	testB	val	testA	testB
MDETR (pretrained)	86.75	89.58	81.41	79.52	84.09	70.62
+SiRi	87.24	89.57	81.83	79.77	84.28	70.98

Table 1. Experiment results of MDETR (large pretraining) with SiRi on RefCOCO and RefCOCO+.

References

1. Jiajun Deng, Zhengyuan Yang, Tianlang Chen, Wengang Zhou, and Houqiang Li. TransVG: End-to-End Visual Grounding with Transformers. *ICCV*, 2021.
2. Zhao Yang and et al. Lavt: Language-aware vision transformer for referring image segmentation. In *CVPR*, 2022.

Fig. 4. Examples of correct comprehension of referring expressions on RefCOCO+.

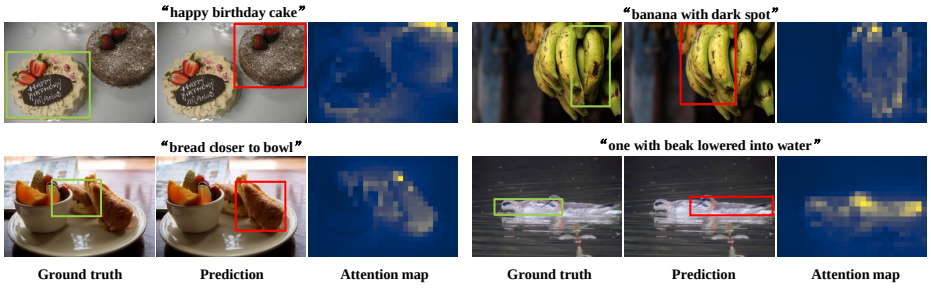


Fig. 5. Failure cases of our model prediction on RefCOCO+.