SafaRi:Adaptive Sequence Transformer for Weakly Supervised Referring Expression Segmentation (Supplementary Material)

Sayan Nag[1,2⋆], Koustava Goswami[2], and Srikrishna Karanam[2]

[1] University of Toronto
sayan.nag@mail.utoronto.ca
[2] Adobe Research
{koustavag,skaranam}@adobe.com

A Additional Related Work

Unsupervised Referring Expression Comprehension (REC). Referring Expression Comprehension (REC) is a grounding task which involves localizing an object present in an image with respect to a textual description of that object [9, 12, 17, 23, 24, 42, 47]. With the introduction of Vision-Language Pretrained (VLP) models [5,8,15,26], it has been possible to develop *Unsupervised* or *Zero-Shot* REC (ZS-REC) methods [11, 31, 33, 43]. CPT [43] colors region proposal boxes and utilizes a captioning model for predicting the colored proposals that are linked to the textual expressions. RedCircle [31] employs visual prompting by drawing circular contours outside the detected object proposals and subsequently ranking them based on the obtained CLIP scores. However, it majorly lacks the understanding of the spatial relationships among the detected proposals.

Referring Video Object Segmentation. Referring Video Object Segmentation (R-VOS) is a cross-modal task with the objective to segment the target object in all video frames, referred by a linguistic description. Existing approaches comprises of (i) *bottom-up* methods where RES algorithms are applied independently at the frame-level [30, 44], (ii) *top-down* methods where a language grounding model selects the best object tracklet from a candidate set of tracklets initially constructed by propagating the detected object masks from key frames [16], and (iii) *query-based* method which introduce a small set of object queries that are conditioned on the referring expression for the target object [38]. However, most of these methods are trained in a fully-supervised manner on the benchmark R-VOS datasets. Conversely, in this work we consider a Zero Shot R-VOS (ZS-R-VOS) task where we apply our model (trained on RES task) in a *zero-shot* manner.

Sequence-to-Sequence (seq2seq) Modeling in Vision Tasks. Seq2seq modeling has displayed immense success in Natural Language Processing (NLP)

⋆ Work done during internship at Adobe Research.

tasks [1, 7, 27, 28, 35, 36]. Taking inspiration from such studies, recent progress have been made to model vision tasks in a seq2seq manner [3,4,19,20,37,41,46]. One of the foremost studies in this domain is Pix2Seq [3] that proposes object detection as a seq2seq modeling task conditioned on the observed pixel inputs. Pix2Seqv2 [4] is an extension of Pix2Seq with an inclusion of instance segmentation and captioning tasks unified into a single shared interface. UniTAB [41] employs a unified seq2seq learning framework which is able to jointly output open-ended text and box representations, facilitating alignment between words and boxes. Obj2Seq [6] takes objects as inputs and outputs human pose estimation into sequence-generated form. Finally, SeqTR [46] proposes to combine visual grounding tasks under a unified framework. Taking inspiration from these studies, we implement RES with a contour prediction approach, however, in a weakly-supervised setting.

B Additional Method Details

B.1 Revisiting SeqTR

SeqTR [46] conceptualized Referring Expression Segmentation (RES) as an auto-regressive point-prediction problem instead of independent pixel classification tasks as employed in [2,12,40]. Unlike [12,45], it further unified Visual Grounding tasks resulting in a simple and efficient multi-task training. In the case of RES, the segmentation mask is serialized into a sequence of N discrete co-ordinate object-contour points given as $\{x_i, y_i\}\ \forall\ i \in \{1, ..., N\}$. During inference, SeqTR predicts the target tokens in an auto-regressive manner ending it with a special [EOS] token. Furthermore, SeqTR involves the use of simple cross-entropy loss for RES task instead of iou or dice losses as used in [2, 12]. Moreover, SeqTR has shown to perform exceedingly well as compared to MDETR [12] in the case of limited annotations [25]. Therefore, taking the above inspiration we develop SafaRi, which outperforms existing baselines in both the fully supervised and weakly supervised scenarios by substantial margins.

B.2 Construction of Contour Points

We followed [25, 46] to construct the sequence of contour points for the RES task. This is done by uniformly sampling P points clockwise along the contour of the mask and subsequently quantizing them into M quantization bins scaled by image height (H) and width (W), using:

$$x_i = round\left(\frac{Mx_i}{W}\right); y_i = round\left(\frac{My_i}{H}\right) \tag{1}$$

B.3 Additional details on Collapse Reduction term of AMCR Loss

For the collapse reduction term, we first compute the total number of non-zero elements in $\mathcal{A}$ given by $n(\mathcal{A})$. This is achieved by computing the sum of the

Algorithm 1 Weak-Supervision with γ-Scheduling

Require: GT data $\mathcal{D}^{GT}$: $\{I, T, M\}$; Remaining Unlabelled data $\mathcal{D}^{GT}_{rem}$: $\{I_{rem}, T_{rem}\}$
Mask Validity Filtering: MVF; Total Scheduling Steps: $\mathcal{S}$; Schedule factor: $\gamma_0 = 0.9$
SafaRi Encoder: $\mathcal{F}^0_\theta$
$\mathcal{L}_{total} \leftarrow \mathcal{F}^0_\theta(I, T, M))$ ▷ Step 1
while $s \leq \mathcal{S}$ **do**
  $\hat{M} \leftarrow \text{MVF}(\mathcal{F}^s_\theta(I_{rem}, T_{rem}))$ ▷ Step 2
  $\{I', T', M'\} \leftarrow \{I, T, M\} + \{\hat{I}, \hat{T}, \hat{M}\}$
  $\mathcal{L}_{\text{SafaRi}} \leftarrow \mathcal{F}^s_\theta(I', T', M')$ ▷ Step 3
  $\gamma_{s+1} \leftarrow \gamma_s + 0.1/\mathcal{S}$
  $s \leftarrow s + 1$
end while

ratio $\frac{\mathcal{A}}{\mathcal{A}+\epsilon}$ where $\epsilon = 0.0001$. Next, we compute the ratio of $n(\mathcal{A})$ to the sum of the mask $\mathcal{M}$ pixel values which gives the total number of non-zero values of the mask for any image b in the batch of size N. We collect the ratios for all the images in the batch generating a distribution $\mathbf{Q}_N$ which in an ideal scenario should entirely overlap with an uniform distribution (target distribution) given by $\mathbf{U}_N$. Therefore, we aim to minimize the Kullback-Liebler (KL) divergence loss between the two distributions.

B.4 Algorithm for Weak-Supervision with γ-Scheduling

The algorithm for Weak-Supervision with γ-Scheduling is given in Algorithm 1.

B.5 SpARC

B.5.1 ZS-REC with SpARC. In our Mask Validity Filtering stage, unlike Partial-RES [25] we **do not use 100% ground truth box annotations**. This is because we consider a more *challenging* and *realistic weakly-supervised* problem setting for RES. Therefore, we propose SpARC, which incorporates red-box visual prompting and spatial awareness components in a pretrained Vision-Language Model (VLM), such as CLIP. SpARC is a zero-shot technique for finding the box corresponding to the object being referred to in the associated linguistic expression. As shown in Figure 1, CLIP takes in the images with objects surrounded by red borders and respective backgrounds being blurred and generates proposal (proposals are detected using an object detector pretrained on a different dataset, see main text) scores based on the text for each of these visual prompts ('A', 'B', 'C'). However, as shown (Figure 1) instead of 'A', it gives highest score for 'C' indicating that spatial understanding is missing in CLIP. The Spatial Reasoning (SR) component in SpARC takes in the bounding boxes and the relation (e.g., in this case: "*left*"). Based on the bounding box information, it generates probability scores (e.g., in this case: 1 for 'A', 0 for both 'B' and 'C') which are then multiplied with the CLIP scores to generate the final combined scores (e.g., as correctly identified, 'A' has the highest score

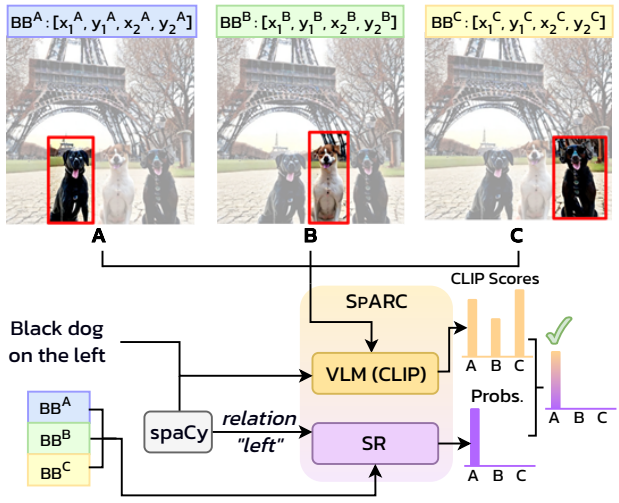

Fig. 1: Example of Zero-Shot Referring Expression Comprehension with SPARC module used for identifying the boundaries (boxes) of the objects of interest used in the Mask Validity Filtering stage. VLM: Vision-Language Model. SR: Spatial Reasoning.

in this case). More details regarding the spatial relations are discussed next. Results on RefCOCO/+/g val and test datasets are also provided in Table 1 demonstrating SPARC significantly outperforms baseline methods on the ZS-REC task.

B.5.2 Spatial Reasoning component in SPARC. In the Spatial Reasoning module of SPARC, we use a rule-based assignment procedure depending on the spatial relationships among different object proposals in the image, as can be identified by their positions, orientations and associations with other objects within the image. We broadly categorize these relationships as:

1. **Position:** We specifically consider the relative position of one object with respect to another object, e.g. "left," "above," etc.
2. **Distance:** We compute the distance between proposals to assess if an object is "near"/"closer" or "far" from the other object.
3. **Size:** We compare the sizes by computing the area of the proposals to determine if an object is "larger" or "smaller" than another object.
4. **Containment/Intersection:** Utilizing IoUs we compute the degree of intersection and assess if one object is fully/partially "inside", or completely "outside". Furthermore, we also check using the location of proposal's centroid if an object lies in "between" two other objects.
5. **Special Relations:** We also analyze some special relations such as "with" by finding close associations using distance metric between two proposals' centroids. Moreover, for cases containing information such as "n o'clock"

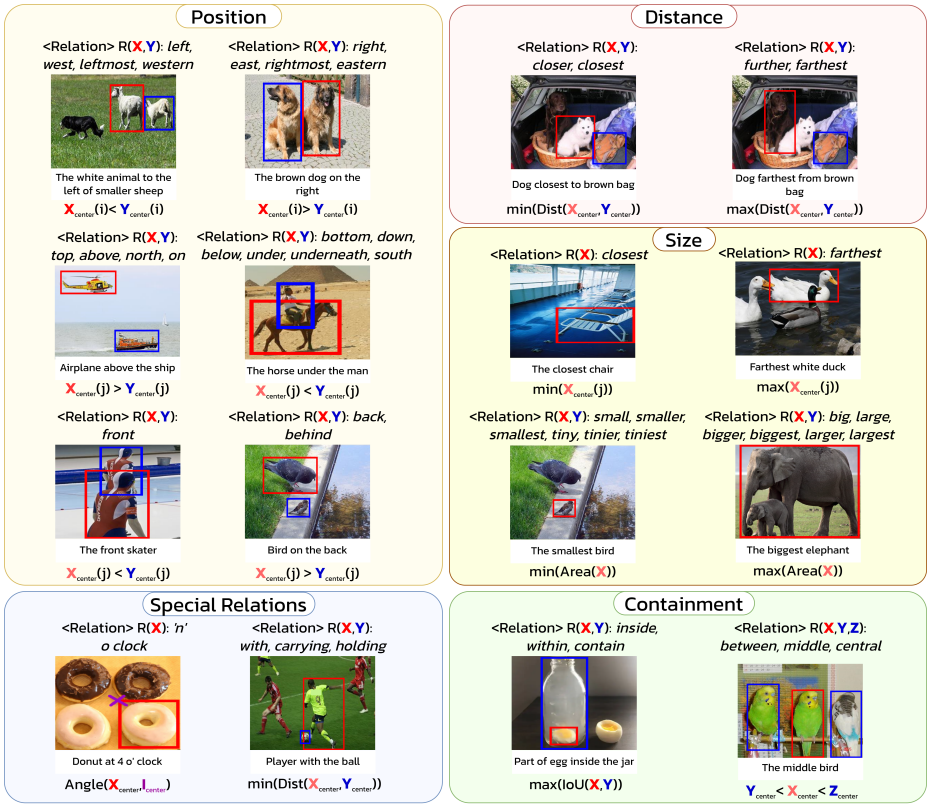

Fig. 2: Examples of Spatial Relations between different objects in images and respective rule-based assignment protocols.

where n refers to a natural number between 1 and 12, we calculate the angle of the proposal's centroid with respect to the center of the image.

We provide such examples of relationships along with their respective rule assignments in Figure 2. Incorporating such information leads to improved performances of SpARC in the MVF stage used for filtering of pseudo-labels.

B.6 Additional Dataset Details

B.6.1 RES. RefCOCO contains 142,209 annotated expressions for 50,000 objects in 19,994 images, and RefCOCO+ comprises of 141,564 expressions for 49,856 objects in 19,992 images. In contrast to RefCOCO, RefCOCO+, does not contain the location words in the referring expressions therefore making it

Method	Eval.	RefCOCO			RefCOCO+			RefCOCOg	
		val	testA	testB	val	testA	testB	val	test
DTWREG [34]	FT	39.2	41.1	37.7	39.2	40.1	38.1	–	–
Pseudo-Q [11]	FT	56.0	58.3	54.1	38.9	45.1	32.1	46.3	47.4
CPT-Blk [43]	ZS	26.9	27.5	27.4	25.4	25.0	27.0	32.1	32.3
CPT-Seg [43]	ZS	32.2	36.1	30.3	31.9	35.2	28.8	36.7	36.5
Red Circle [31]	ZS	49.8	58.6	39.9	55.3	63.9	45.4	59.4	58.9
SPARC (Ours)	**ZS**	**53.5**	**59.0**	**55.6**	**55.8**	**65.1**	**52.8**	**68.8**	**67.5**
$\Delta_{\text{SPARC - RedCircle}}$	ZS	3.7 ↑	0.4 ↑	15.7 ↑	0.5 ↑	1.2 ↑	7.4 ↑	9.4 ↑	8.6 ↑

Table 1: Comparison of SOTA methods on ZS-REC task. SPARC outperforms SOTA baselines by significant margins across all the datasets.

Method	RefCOCO@val	RefCOCO+@val	RefCOCOg@val
Strudel et al. (2022) [32]	25.95	22.62	23.41
Kim et al. (2023) [13]	34.76	28.48	28.87
Lee et al. (2023) [14]	31.06	31.28	32.88
Liu et al. (2023) [18]	31.17	30.90	36.19
SAFARI-10 (Ours)	**64.02**	**52.98**	**52.91**

Table 2: SAFARI versus other text-based WSRES methods.

Method	RefCOCO@val	RefCOCO+@val	RefCOCOg@val
CLIP-DIY [39]	22.15	20.88	20.65
SAFARI-10 (Ours)	**64.02**	**52.98**	**52.91**

Table 3: Comparison of SAFARIwith CLIP-DIY.

Method	RefCOCO@val	RefCOCO+@val	RefCOCOg@val
Grounded-SAM [29]	75.98	70.43	69.17
SAFARI(Ours)	**77.21**	**70.78**	**70.48**

Table 4: SAFARI vs Grounded-SAM on fully-supervised RES.

Mask Validity Filtering	RefCOCO@val
✗	62.68
✓	**67.04**

Table 5: Effect of Mask Validity Filtering (MVF) on the final performance of SAFARI on RefCOCO@val dataset for 30% mask annotations. Without MVF, the pseudo-masks are not validated and performance drops significantly.

Cross-Attention Parameter (β)	RefCOCO@val
β = 0.25	63.59
β = 0.50	64.42
β = 0.75	62.11
Learnable β	**67.04**

Table 6: Ablation on the gated cross-attention parameter (β) on RefCOCO@val dataset for 30% mask annotations. Learnable gating performs as compared to a fixed value.

more challenging as a dataset. RefCOCOg has 85,474 referring expressions for 54,822 objects in 26,711 images. Here, the expressions are florid, descriptive and complex with an average of 8.4 words per text (as compared to 3.5 words of RefCOCO and RefCOCO+). Following [25, 46], we resort to using the UMD split [21] for RefCOCOg.

DSC thresholding parameter (τ)	RefCOCO val
τ = 1.0	62.68
τ = 0.5	63.31
τ = 0.05	64.55
τ = 0.1	**67.04**

Table 7: Effect of DSC threshold parameter (τ) on the final performance of SafaRi on RefCOCO@val dataset for 30% mask annotations. With τ = 0.1 we get the best performance. Increasing τ degrades performance.

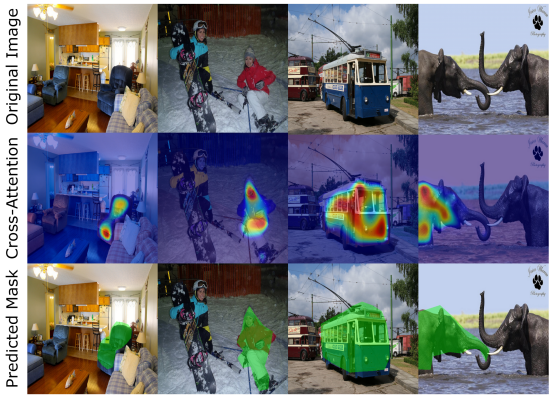

Table 8: Extended qualitative examples of cross-attention maps and corresponding predictions showing strong cross-modal alignment learned by SafaRi.

Fig. 3: Extended examples of masks with increasing WSRES bootstrapping runs (steps) for 10% annotations. We observe significant improvements in grounding capabilities of SafaRi with an increase in semi-supervised retraining steps depicting the importance of the retraining stage.

B.6.2 ZS-R-VOS. Ref-DAVIS17 has been built upon DAVIS17 dataset [22] by incorporating an associated referring expression for a particular object present in each video frame. Specifically, there are 90 videos with 1,544 referring expressions for 205 objects present in this dataset. Training and validation sets contain 60 and 30 videos respectively. JHMDB-Sentences is created by adding language annotations on the original JHMDB dataset [10]. In total, JHMDB-Sentences has 928 videos each associated with a referring expression.

B.7 Additional Implementation Details

Following [25, 46], we resized images to 640 × 640, and set maximum token length to 15 for RefCOCO/RefCOCO+ and 20 for RefCOCOg. We uniformly

Fig. 4: Extended examples of predicted masks with varying label-rates. It is evident that with increasing mask annotation percentage (%), the quality of predicted masks improves.

sampled contour points from the ground truth masks generating 18 points for RefCOCO/RefCOCO+, and 15 for RefCOCOg.

C Additional Results

C.1 ZS-REC based MVF using SpARC. MVF stage is responsible for filtering pseudo-masks to be used in the semi-supervised retraining stage. Thus, we evaluate the performance of SpARC module in ZS-REC task on RefCOCO/+/g val and test sets in Table 1. We notice significant improvements across all datasets surpassing existing SOTA methods, displaying tremendous generalization capabilities.

C.2 Comparison with Text-only WSRES methods. Notably, SafaRi substantially outperforms text-only weakly-supervised (WS) RES methods (Tab 2) indicating the benefit of training with text plus few mask/box annotations (e.g., 10 %) rather than solely relying on text. However, a drawback is the use of these expensive grounding information in training (albeit few).

C.3 Comparison with CLIP-DIY. We report results with CLIP-DIY in Tab 3 where SafaRi with just 10% annotations significantly outperforms CLIP-DIY. A possible reason is that RES tasks require spatial reasoning capabilities which is a limitation in CLIP models. Moreover, CLIP text encoder has been trained with simple captions like "A photo of an <object>" and encoding complex referring texts (as in RefCOCO) is challenging. Therefore, it is important to have a weakly-supervised training paradigm (with as low as 10% annotations) to obtain a substantial performance on challenging RES tasks.

C.4 Comparison with Grounded-SAM on *fully-supervised* RES. Grounded-SAM (G-SAM) uses box outputs from G-DINO as prompts for SAM to predict segmentation masks [29]. Notably, as the reviewer acknowledges, a direct comparison is infeasible due to differences in tasks (our focus is weakly-supervised

Fig. 5: Extended examples of predicted masks on RefDAVIS17 and JHMDB datasets in a zero-shot setting with SafaRi trained with 30% annations.

RES). Also, unlike ours, these models are extensively pretrained on large-scale datasets. Since G-SAM did not report results on fully supervised RES, we reimplement it (Tab 4).

D Additional Ablations

D.1 Impact of Mask Validity Filtering (MVF). We assess the effect of Mask Validity Filtering approach in our WSRES pipeline in Table 5. We observe that in the absence of the MVF stage, the generated pseudo-labels are not validated which hurts the overall prediction performance of the model. Conversely, in the presence of MVF, the mIoU values on RefCOCO@val set are found to improve by 4.36 mIoU, displaying the effectiveness of the MVF stage.

D.2 Impact of Gated Cross-Attention parameter (β). We investigate the importance of gated cross-attention modules with learnable gating scalar parameter (β) on the final performance of SafaRi on RefCOCO@val set with 30% mask annotations in Table 6. We notice that a fixed value of β leads to a significant drop in mIoU value as compared to when it is being kept learnable.

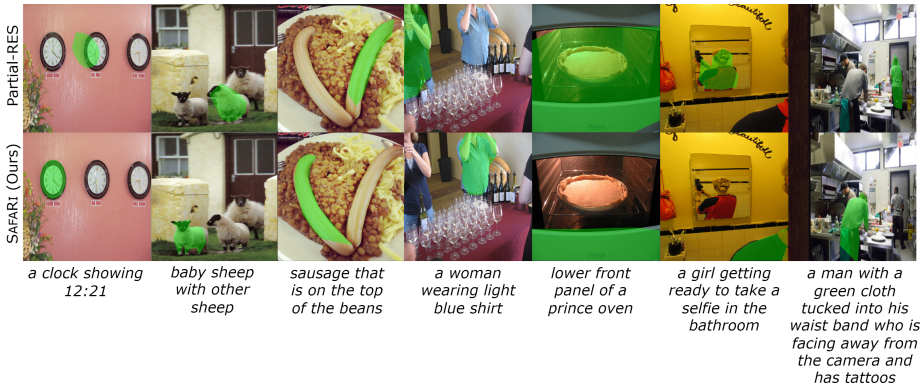

Fig. 6: Comparison of Partial-RES and SafaRi on the WSRES task on RefCOCO/+/g validation sets when trained using 30% mask annotations. SafaRi possesses excellent grounding capabilities under challenging scenarios with complex linguistic expressions. Partial-RES is found to fail in such cases.

Fig. 7: Limitations of our method. Tiny and vastly hindered objects, especially under low-light and blurry conditions are not distinctly localized by SafaRi.

D.3 Impact of DSC thresholding parameter (τ). We study the effect of DSC thresholding patrameter (τ) on the model performance on RefCOCO@val set with 30% mask annotations in Table 7. Results indicate that τ = 0.10 is an optimal threshold value.

E Additional Qualitative Assessment

In Figures 8, 3, 4, and 5, we provide extended examples of the ones analyzed in Section 4.4. We further provided qualitative comparisons between Partial-RES [25] and SafaRi and also display failure cases of SafaRi. Additionally, we demonstrate qualitative examples for the ZS-REC task using our SpARC approach as compared to RedCircle [31].

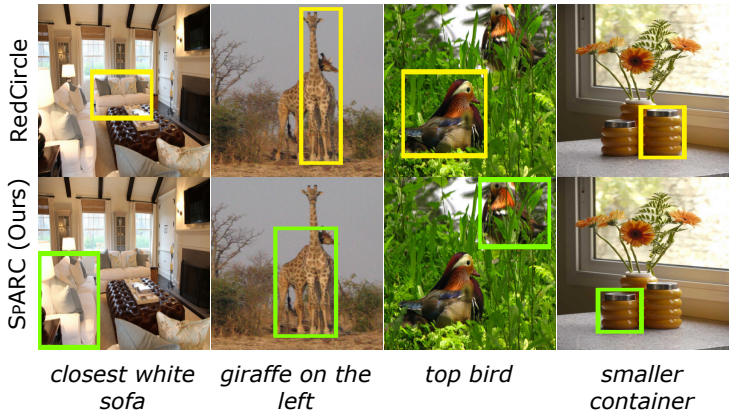

Fig. 8: Qualitative examples of detected bounding boxes using our proposed SpARC as compared to RedCircle on the ZS-REC task with RefCOCO/+/g validation sets. The spatial reasoning module paired with CLIP aids in detecting the correct bounding boxes which is not possible with just a CLIP based approach such as RedCircle which does not encode the spatial relationships among objects in the image.

E.1 Comparisons of SafaRi with Partial-RES. The novel components of our framework, such as the cross-modal attention based fused feature extractors, the AMCR loss and the bootstrapping with MVF module not only makes our approach unique but also equips SafaRi with excellent grounding capabilities, especially in challenging cases with complex referring expressions as shown in Figure 6. It is worth noting that although (unlike our approach) Partial-RES [25] uses 100% box annotations, it fails to focus on the referred objects in these complex situations as depicted in Figure 6. This once again highlights the importance of having cross-attention based feature fusion along with AMCR in the X-FACt module in our framework.

E.2 Limitations of SafaRi. Harnessing the power of cross-modal fusion, SafaRi learns impressive cross-modal representation which assists in generating excellent high-quality segmentation masks. However, there exists some cases where SafaRi fails to identify tiny and hindered objects, especially in cluttered environments, under low-light and blurry conditions. We show such examples in Figure 7. For example, in the two cases of *'a man holding a tv remote control'* and *'girl with green hair tie'*, the respective referred objects *'tv remote control'* and *'hair tie'* are extremely tiny and barely visible. Therefore, SafaRi is unable learn the correct linguistic associations and thereby fails to ground them. In the second example, *'a man with a white shirt behind a table'* is immensely hindered and therefore hard to focus on. In the third example, sunlight falling on the *'sliced bananas'* makes it extremely hard to recognize, even in human eyes. Finally, the *'person in raincoat'* is also not clearly distinguishable because

of low-light conditions and transparent nature of the raincoat which can only be identified upon zooming in by few magnitudes. However, higher-resolution images may be more helpful in addressing such intricate scenarios, which can be considered as a potential future work.

E.3 ZS-REC visualizations with SPARC vs RedCircle. In Figure 8 we demonstrate qualitative comparisons between RedCircle [31] and our proposed SPARC approach on the ZS-REC task. It is evident that spatial reasoning module in conjunction with CLIP leads to identification of correct bounding boxes in cases where RedCircle fails. It is worth noting that RedCircle [31] is a visual prompting approach based on CLIP which does not possess spatial reasoning capabilities.

References

1. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. In: NeurIPS (2020) 2
2. Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., Zagoruyko, S.: End-to-end object detection with transformers. In: ECCV. pp. 213–229 (2020) 2
3. Chen, T., Saxena, S., Li, L., Fleet, D.J., Hinton, G.: Pix2seq: A language modeling framework for object detection. arXiv preprint arXiv:2109.10852 (2021) 2
4. Chen, T., Saxena, S., Li, L., Lin, T.Y., Fleet, D.J., Hinton, G.: A unified sequence interface for vision tasks. In: NeurIPS (2022), https://openreview.net/forum?id=tjFaqsSK2I3 2
5. Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Universal image-text representation learning. In: ECCV. pp. 104–120 (2020) 1
6. Chen, Z., Zhu, Y., Li, Z., Yang, F., Li, W., Wang, H., Zhao, C., Wu, L., Zhao, R., Wang, J., et al.: Obj2seq: Formatting objects as sequences with class prompt for visual tasks. NeurIPS **35**, 2494–2506 (2022) 2
7. Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning phrase representations using RNN encoder–decoder for statistical machine translation. In: EMNLP. pp. 1724–1734 (2014). https://doi.org/10.3115/v1/D14-1179, https://aclanthology.org/D14-1179 2
8. Chowdhury, S., Nag, S., Manocha, D.: Apollo: Unified adapter and prompt learning for vision language models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 10173–10187 (2023) 1
9. Hong, R., Liu, D., Mo, X., He, X., Zhang, H.: Learning to compose and reason with language tree structures for visual grounding. IEEE TPAMI (2019) 1
10. Jhuang, H., Gall, J., Zuffi, S., Schmid, C., Black, M.J.: Towards understanding action recognition. In: ICCV. pp. 3192–3199 (2013) 7
11. Jiang, H., Lin, Y., Han, D., Song, S., Huang, G.: Pseudo-q: Generating pseudo language queries for visual grounding. In: CVPR. pp. 15513–15523 (2022) 1, 6
12. Kamath, A., Singh, M., LeCun, Y., Synnaeve, G., Misra, I., Carion, N.: Mdetr-modulated detection for end-to-end multi-modal understanding. In: ICCV. pp. 1780–1790 (2021) 1, 2
13. Kim, D., Kim, N., Lan, C., Kwak, S.: Shatter and gather: Learning referring image segmentation with text supervision. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15547–15557 (2023) 6

14. Lee, J., Lee, S., Nam, J., Yu, S., Do, J., Taghavi, T.: Weakly supervised referring image segmentation with intra-chunk and inter-chunk consistency. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21870–21881 (2023) 6
15. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. In: NeurIPS (2021) 1
16. Liang, C., Wu, Y., Zhou, T., Wang, W., Yang, Z., Wei, Y., Yang, Y.: Rethinking cross-modal interaction from a top-down perspective for referring video object segmentation. arXiv preprint arXiv:2106.01061 (2021) 1
17. Liao, Y., Liu, S., Li, G., Wang, F., Chen, Y., Qian, C., Li, B.: A real-time cross-modality correlation filtering method for referring expression comprehension. In: CVPR. pp. 10880–10889 (2020) 1
18. Liu, F., Liu, Y., Kong, Y., Xu, K., Zhang, L., Yin, B., Hancke, G., Lau, R.: Referring image segmentation using text supervision. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22124–22134 (2023) 6
19. Liu, J., Ding, H., Cai, Z., Zhang, Y., Satzoda, R.K., Mahadevan, V., Manmatha, R.: Polyformer: Referring image segmentation as sequential polygon generation. In: CVPR. pp. 18653–18663 (2023) 2
20. Lu, J., Clark, C., Zellers, R., Mottaghi, R., Kembhavi, A.: Unified-IO: A unified model for vision, language, and multi-modal tasks. arXiv preprint arXiv:2206.08916 (2022) 2
21. Nagaraja, V.K., Morariu, V.I., Davis, L.S.: Modeling context between objects for referring expression understanding. In: ECCV. pp. 792–807 (2016) 6
22. Pont-Tuset, J., Perazzi, F., Caelles, S., Arbeláez, P., Sorkine-Hornung, A., Van Gool, L.: The 2017 davis challenge on video object segmentation. arXiv preprint arXiv:1704.00675 (2017) 7
23. Pramanick, S., Han, G., Hou, R., Nag, S., Lim, S.N., Ballas, N., Wang, Q., Chellappa, R., Almahairi, A.: Jack of all tasks master of many: Designing general-purpose coarse-to-fine vision-language model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14076–14088 (2024) 1
24. Pramanick, S., Jing, L., Nag, S., Zhu, J., Shah, H., LeCun, Y., Chellappa, R.: Volta: Vision-language transformer with weakly-supervised local-feature alignment. arXiv preprint arXiv:2210.04135 (2022) 1
25. Qu, M., Wu, Y., Wei, Y., Liu, W., Liang, X., Zhao, Y.: Learning to segment every referring object point by point. In: CVPR. pp. 3021–3030 (2023) 2, 3, 6, 7, 10, 11
26. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: ICML. pp. 8748–8763 (2021), https://proceedings.mlr.press/v139/radford21a.html 1
27. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019) 2
28. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer. JMLR (2020) 2
29. Ren, T., Liu, S., Zeng, A., Lin, J., Li, K., Cao, H., Chen, J., Huang, X., Chen, Y., Yan, F., et al.: Grounded sam: Assembling open-world models for diverse visual tasks. arXiv preprint arXiv:2401.14159 (2024) 6, 8
30. Seo, S., Lee, J.Y., Han, B.: Urvos: Unified referring video object segmentation network with a large-scale benchmark. In: ECCV. pp. 208–223 (2020) 1

31. Shtedritski, A., Rupprecht, C., Vedaldi, A.: What does clip know about a red circle? visual prompt engineering for vlms. arXiv preprint arXiv:2304.06712 (2023) 1, 6, 10, 12
32. Strudel, R., Laptev, I., Schmid, C.: Weakly-supervised segmentation of referring expressions. arXiv preprint arXiv:2205.04725 (2022) 6
33. Subramanian, S., Merrill, W., Darrell, T., Gardner, M., Singh, S., Rohrbach, A.: Reclip: A strong zero-shot baseline for referring expression comprehension. arXiv preprint arXiv:2204.05991 (2022) 1
34. Sun, M., Xiao, J., Lim, E.G., Liu, S., Goulermas, J.Y.: Discriminative triad matching and reconstruction for weakly referring expression grounding. IEEE TPAMI pp. 4189–4195 (2021) 6
35. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to sequence learning with neural networks. In: NeurIPS (2014) 2
36. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023) 2
37. Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., Yang, H.: OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In: ICML. pp. 23318–23340 (2022), https://proceedings.mlr.press/v162/wang22al.html 2
38. Wu, J., Jiang, Y., Sun, P., Yuan, Z., Luo, P.: Language as queries for referring video object segmentation. In: CVPR. pp. 4974–4984 (2022) 1
39. Wysoczańska, et al.: Clip-diy: Clip dense inference yields open-vocabulary semantic segmentation for-free. In: WACV (2024) 6
40. Yang, S., Xia, M., Li, G., Zhou, H.Y., Yu, Y.: Bottom-up shift and reasoning for referring image segmentation. In: CVPR. pp. 11266–11275 (2021) 2
41. Yang, Z., Gan, Z., Wang, J., Hu, X., Ahmed, F., Liu, Z., Lu, Y., Wang, L.: Unitab: Unifying text and box outputs for grounded vision-language modeling. In: ECCV. pp. 521–539 (2022) 2
42. Yang, Z., Gong, B., Wang, L., Huang, W., Yu, D., Luo, J.: A fast and accurate one-stage approach to visual grounding. In: ICCV. pp. 4683–4693 (2019) 1
43. Yao, Y., Zhang, A., Zhang, Z., Liu, Z., Chua, T.S., Sun, M.: Cpt: Colorful prompt tuning for pre-trained vision-language models. arXiv preprint arXiv:2109.11797 (2021) 1, 6
44. Ye, L., Rochan, M., Liu, Z., Wang, Y.: Cross-modal self-attention network for referring image segmentation. In: CVPR. pp. 10502–10511 (2019) 1
45. Yu, L., Lin, Z., Shen, X., Yang, J., Lu, X., Bansal, M., Berg, T.L.: Mattnet: Modular attention network for referring expression comprehension. In: CVPR (June 2018) 2
46. Zhu, C., Zhou, Y., Shen, Y., Luo, G., Pan, X., Lin, M., Chen, C., Cao, L., Sun, X., Ji, R.: Seqtr: A simple yet universal network for visual grounding. In: ECCV. pp. 598–615 (2022) 2, 6, 7
47. Zhuang, B., Wu, Q., Shen, C., Reid, I., Van Den Hengel, A.: Parallel attention: A unified framework for visual object discovery through dialogs and queries. In: CVPR. pp. 4252–4261 (2018) 1