SAFARI: Adaptive Sequence Transformer for Weakly Supervised Referring Expression Segmentation

Sayan Nag[1], Koustava Goswami[2], Srikrishna Karanam[2]
[1]University of Toronto, [2]Adobe Research

Overview

1. Existing Referring Expression Segmentation (RES) approaches need large-scale mask annotations which are expensive to obtain and requires extensive manual interventions.

2. These methods cannot generalize well to unseen scenarios/objects.

3. They also predominantly consider segmentation as an independent pixel classification task neglecting structural information of the objects.

4. To address the aforementioned issues, we propose SafaRi, a weakly-supervised bootstrapping architecture for RES with several new algorithmic innovations.

Contributions

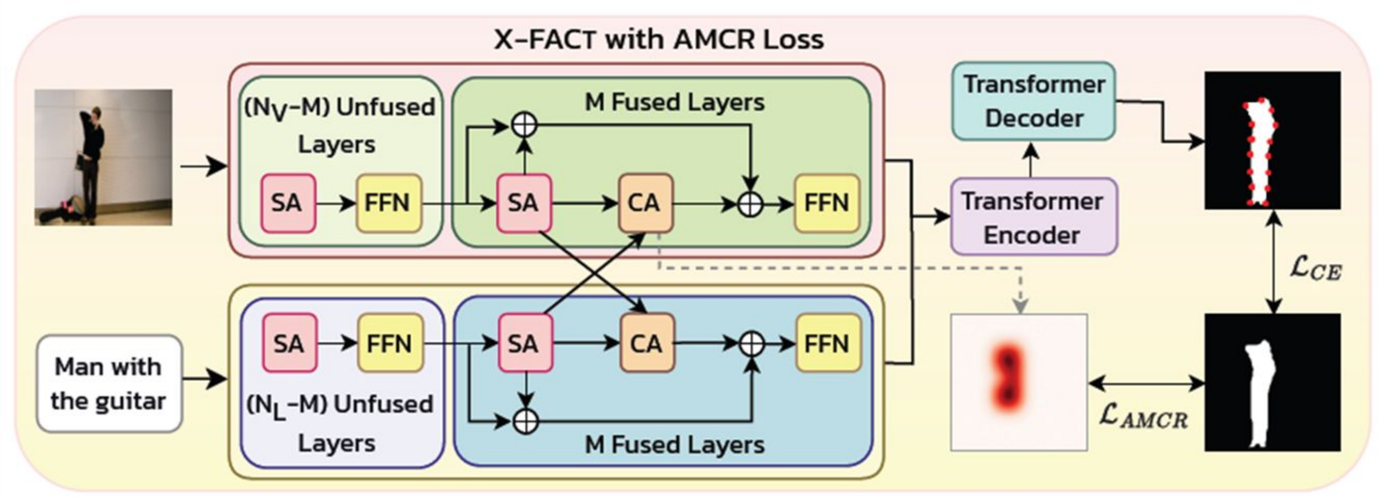
- To the best of our knowledge, **ours is the first** to consider an **accurate representation** of Weakly-Supervised Referring Expression (WS-RES) task by considering a novel, more practical and challenging scenario with limited box and mask annotations where **box % equals mask %**.

- To enable principled training of models in such low-annotation settings, improve **image-text region-level alignment**, and further enhance **spatial localization** of the target object in the image, we propose Cross-modal Fusion with Attention Consistency module **X-FACT**.

- For **automatic pseudo-labeling** of unlabeled samples, we introduce a novel Mask Validity Filtering routine based on a spatially aware zero-shot proposal scoring approach **SPARC**.

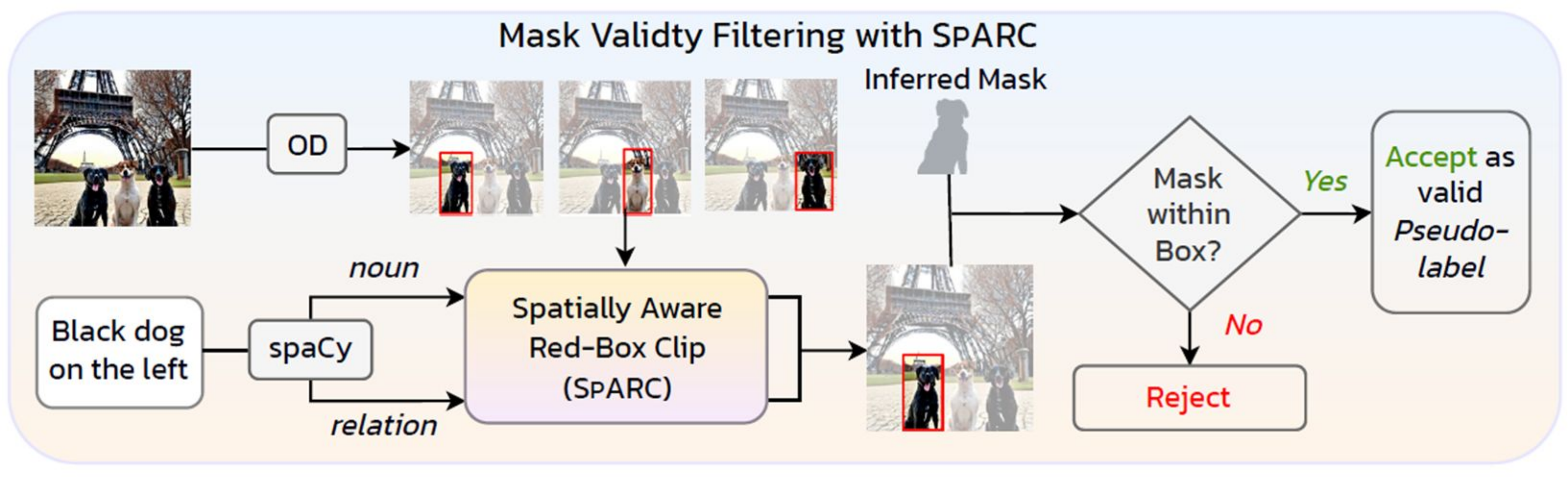
- Extensive experiments demonstrate the efficacy of SafaRi as it **significantly outperforms** baseline models on RES benchmarks. SafaRi also demonstrates **strong generalization capabilities** when evaluated on an unseen referring video object segmentation task in a **zero-shot manner**.

Architecture Overview: X-FACT with AMCR



Overview: We introduce X-FACT, composed of normalized gated cross-attention based Fused Feature Extractors and Attention Consistency Mask Regularization (AMCR) for enhancing cross-modal synergy and spatial localization of target objects. The fused output is subsequently fed to Sequence Transformer for prediction of contour points.

Architecture Overview: Mask Validity Filtering with SPARC

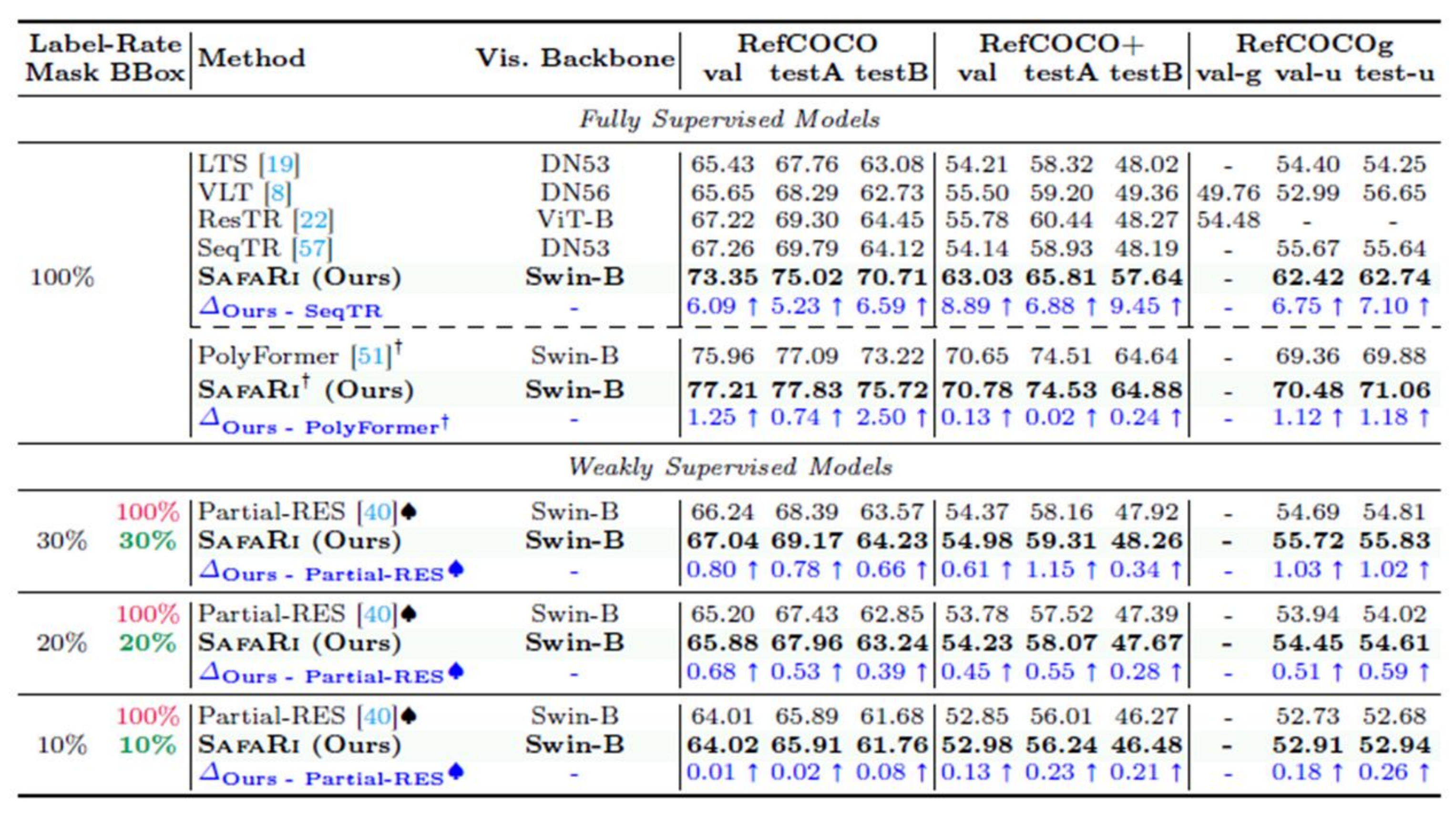

Mask Validty Filtering with SPARC
OD
Inferred Mask
Mask within Box?
Yes
Accept as valid Pseudo-label
No
Reject
noun
Black dog on the left
spaCy
Spatially Aware Red-Box Clip (SPARC)
relation

Overview: We design Mask Validity Filtering (MVF) strategy for choosing valid pseudo-masks using SPARC module which is a Zero-Shot REC approach with spatial reasoning capabilities.

Datasets and Tasks

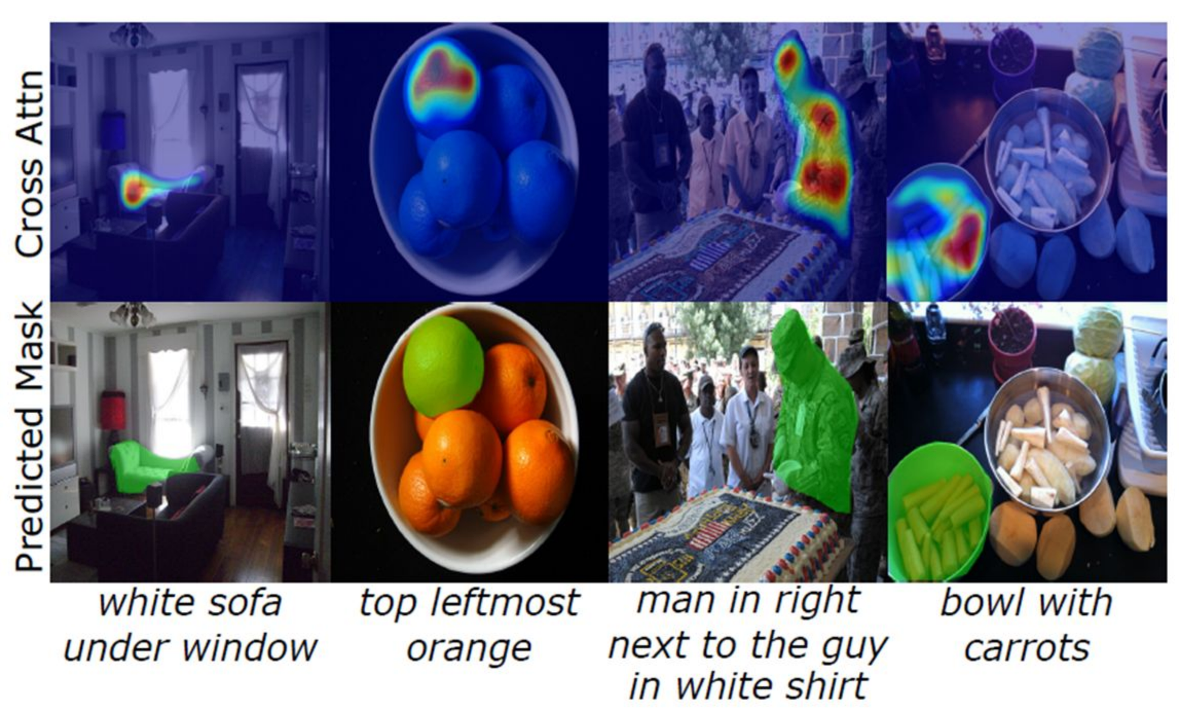
1. RES on Images:
    a. RefCOCO: 142,209 annotated expressions for 50,000 objects in 19,994 images
    b. RefCOCO+: 141,564 expressions for 49,856 objects in 19,992 images
    c. RefCOCOg: 85,474 referring expressions for 131 132 54,822 objects in 26,711 images
2. RES on Videos (Zero- Shot Transfer):
    a. Ref-DAVIS17: 90 videos with 1,544 referring expressions for 205 objects
    b. JHMDB-Sentences: 928 videos each associated with a referring expression

Results on RefCOCO/+/g



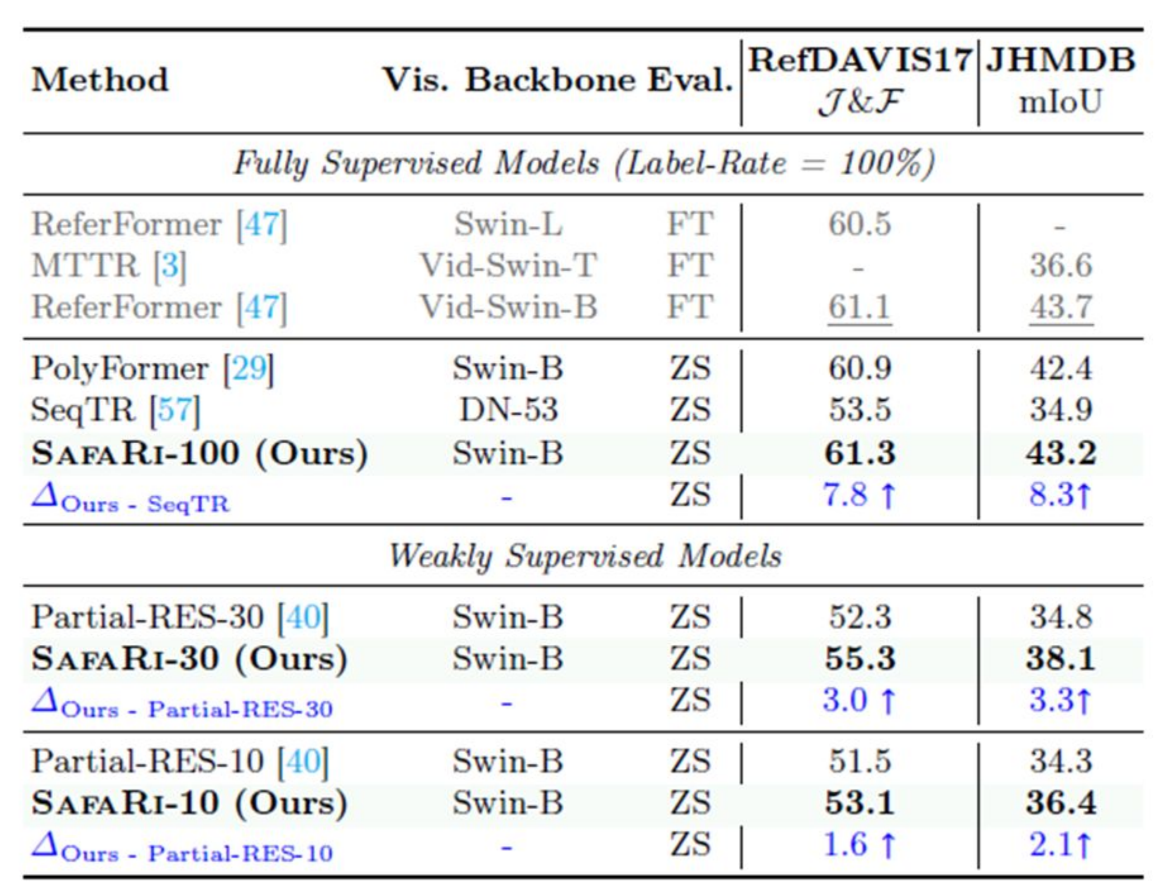
Label-Rate		Method	Vis. Backbone	RefCOCO			RefCOCO+			RefCOCOg		
Mask	BBox			val	testA	testB	val	testA	testB	val-g	val-u	test-u
		Fully Supervised Models										
100%		LTS [19]	DN53	65.43	67.76	63.08	54.21	58.32	48.02	-	54.40	54.25
		VLT [8]	DN56	65.65	68.29	62.73	55.50	59.20	49.36	49.76	52.99	56.65
		ResTR [22]	ViT-B	67.22	69.30	64.45	55.78	60.44	48.27	54.48	-	-
		SeqTR [57]	DN53	67.26	69.79	64.12	54.14	58.93	48.19	-	55.67	55.64
		SaFaRi (Ours)	**Swin-B**	**73.35**	**75.02**	**70.71**	**63.03**	**65.81**	**57.64**	-	**62.42**	**62.74**
		$\Delta_{\text{Ours - SeqTR}}$	-	6.09 ↑	5.23 ↑	6.59 ↑	8.89 ↑	6.88 ↑	9.45 ↑	-	6.75 ↑	7.10 ↑
		PolyFormer [51]†	Swin-B	75.96	77.09	73.22	70.65	74.51	64.64	-	69.36	69.88
		SaFaRi† (Ours)	**Swin-B**	**77.21**	**77.83**	**75.72**	**70.78**	**74.53**	**64.88**	-	**70.48**	**71.06**
		$\Delta_{\text{Ours - PolyFormer}^{\dagger}}$	-	1.25 ↑	0.74 ↑	2.50 ↑	0.13 ↑	0.02 ↑	0.24 ↑	-	1.12 ↑	1.18 ↑
		Weakly Supervised Models										
30%	100%	Partial-RES [40]♠	Swin-B	66.24	68.39	63.57	54.37	58.16	47.92	-	54.69	54.81
	30%	**SaFaRi (Ours)**	**Swin-B**	**67.04**	**69.17**	**64.23**	**54.98**	**59.31**	**48.26**	-	**55.72**	**55.83**
		$\Delta_{\text{Ours - Partial-RES}^{\spadesuit}}$	-	0.80 ↑	0.78 ↑	0.66 ↑	0.61 ↑	1.15 ↑	0.34 ↑	-	1.03 ↑	1.02 ↑
20%	100%	Partial-RES [40]♠	Swin-B	65.20	67.43	62.85	53.78	57.52	47.39	-	53.94	54.02
	20%	**SaFaRi (Ours)**	**Swin-B**	**65.88**	**67.96**	**63.24**	**54.23**	**58.07**	**47.67**	-	**54.45**	**54.61**
		$\Delta_{\text{Ours - Partial-RES}^{\spadesuit}}$	-	0.68 ↑	0.53 ↑	0.39 ↑	0.45 ↑	0.55 ↑	0.28 ↑	-	0.51 ↑	0.59 ↑
10%	100%	Partial-RES [40]♠	Swin-B	64.01	65.89	61.68	52.85	56.01	46.27	-	52.73	52.68
	10%	**SaFaRi (Ours)**	**Swin-B**	**64.02**	**65.91**	**61.76**	**52.98**	**56.24**	**46.48**	-	**52.91**	**52.94**
		$\Delta_{\text{Ours - Partial-RES}^{\spadesuit}}$	-	0.01 ↑	0.02 ↑	0.08 ↑	0.13 ↑	0.23 ↑	0.21 ↑	-	0.18 ↑	0.26 ↑

Results on RefCOCO/+/g



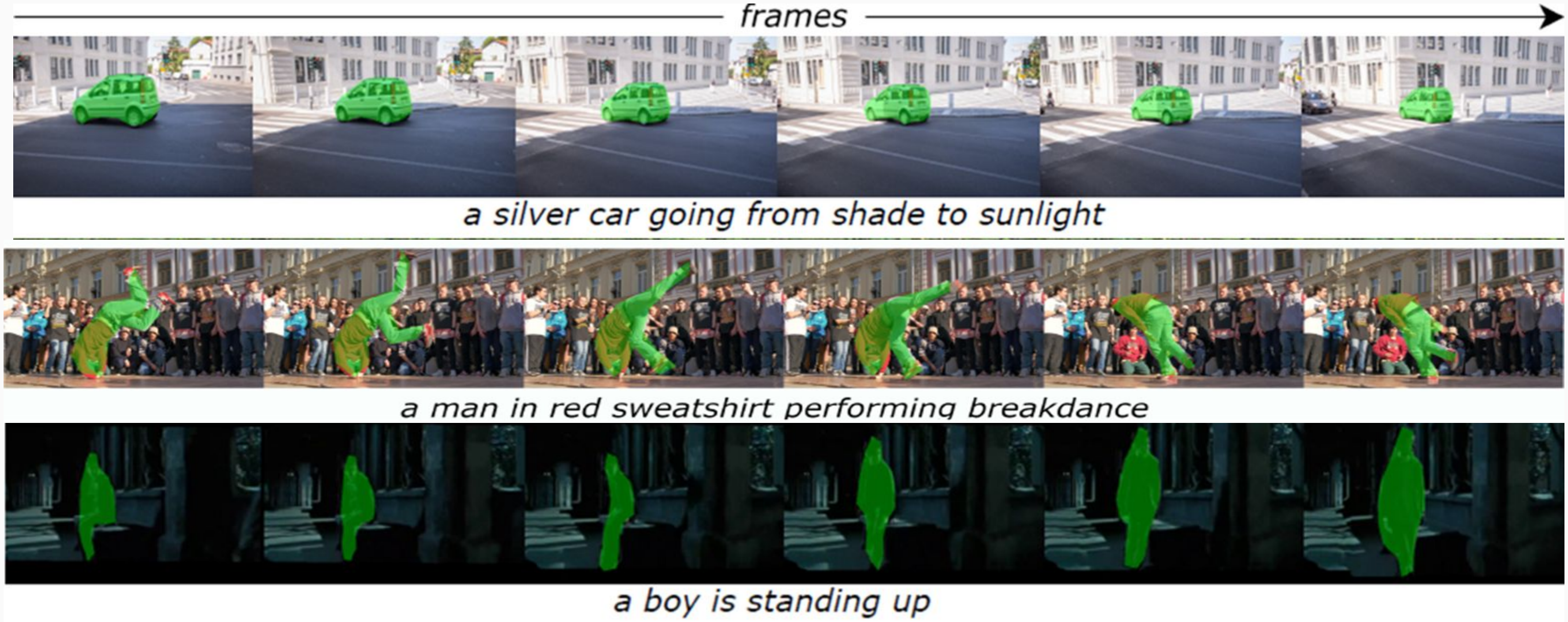
- Cross-modal attention maps attend to different objects in the images, guided by the referring texts.
- The cross-attention scores between the image and the associated expression are extracted and bilinearly interpolated to match the image dimension and superimposed on the original image

Zero-Shot Results on Ref-DAVIS17 and J-HMDB



Method	Vis. Backbone	Eval.	RefDAVIS17	JHMDB
			$\mathcal{J}\&\mathcal{F}$	mIoU
Fully Supervised Models (Label-Rate = 100%)				
ReferFormer [47]	Swin-L	FT	60.5	-
MTTR [3]	Vid-Swin-T	FT	-	36.6
ReferFormer [47]	Vid-Swin-B	FT	61.1	43.7
PolyFormer [29]	Swin-B	ZS	60.9	42.4
SeqTR [57]	DN-53	ZS	53.5	34.9
SafaRi-100 (Ours)	Swin-B	ZS	**61.3**	**43.2**
$\Delta_{\text{Ours - SeqTR}}$	-	ZS	7.8 ↑	8.3↑
Weakly Supervised Models				
Partial-RES-30 [40]	Swin-B	ZS	52.3	34.8
SafaRi-30 (Ours)	Swin-B	ZS	**55.3**	**38.1**
$\Delta_{\text{Ours - Partial-RES-30}}$	-	ZS	3.0 ↑	3.3↑
Partial-RES-10 [40]	Swin-B	ZS	51.5	34.3
SafaRi-10 (Ours)	Swin-B	ZS	**53.1**	**36.4**
$\Delta_{\text{Ours - Partial-RES-10}}$	-	ZS	1.6 ↑	2.1↑

Zero-Shot Results on Ref-DAVIS17 and J-HMDB



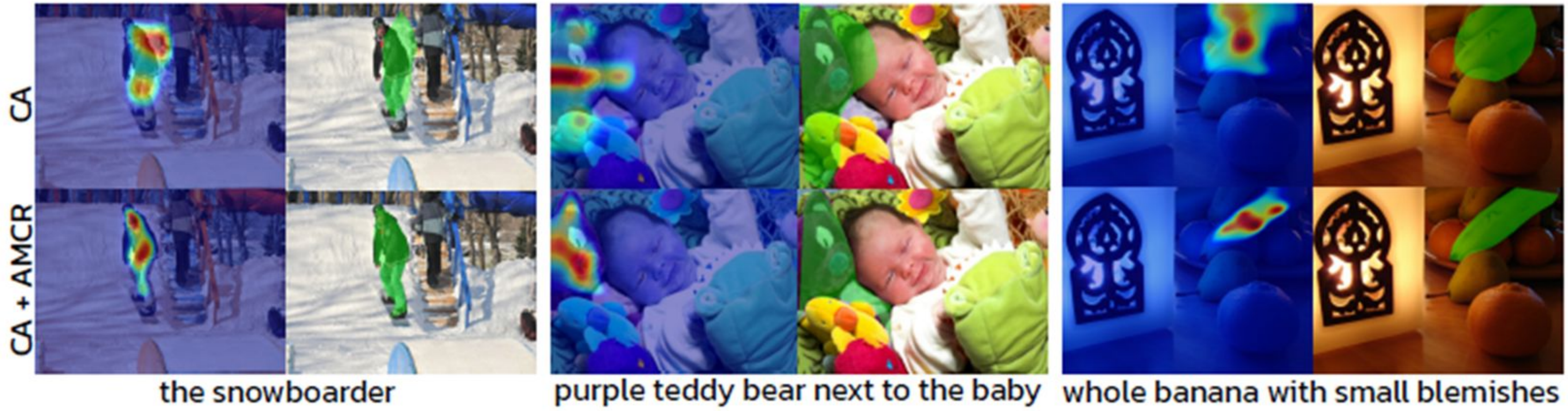
a silver car going from shade to sunlight

a man in red sweatshirt performing breakdance

a boy is standing up

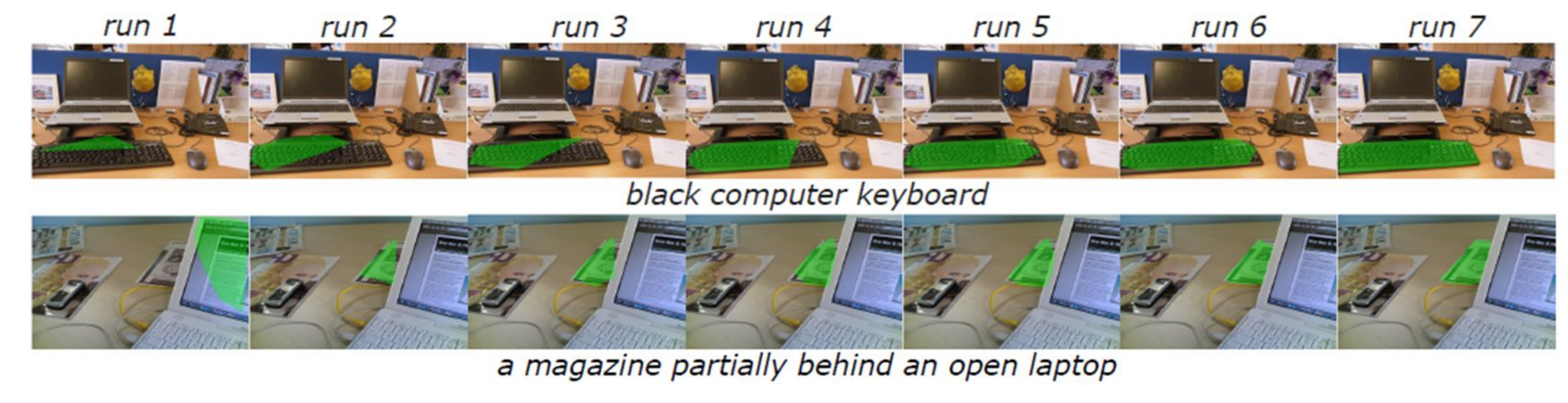
- High-quality segmentation masks generated by SAFARI for the zero-shot task without finetuning
- No temporal information is encoded, still performs well on video datasets

Impact of AMCR


CA
CA + AMCR
the snowboarder
purple teddy bear next to the baby
whole banana with small blemishes

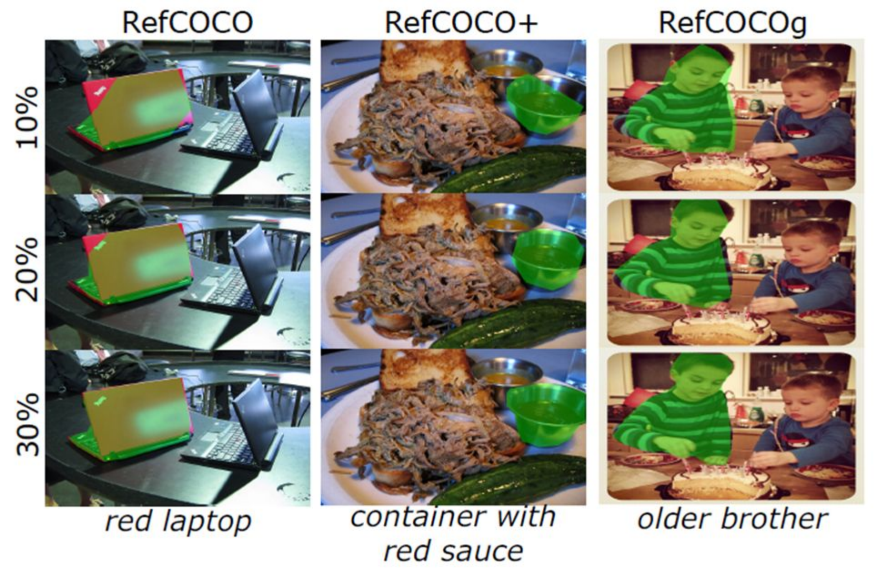
Label-Rate	Cross-Attn.	AMCR	RefCOCO val	RefCOCO testA	RefCOCO testB
30	✗	✗	60.91	64.06	57.73
	✓	✗	64.72	67.55	60.91
	✓	✓	**67.04**	**69.17**	**64.23**
10	✗	✗	54.06	55.64	52.61
	✓	✗	60.11	62.18	57.95
	✓	✓	**64.02**	**65.91**	**61.76**

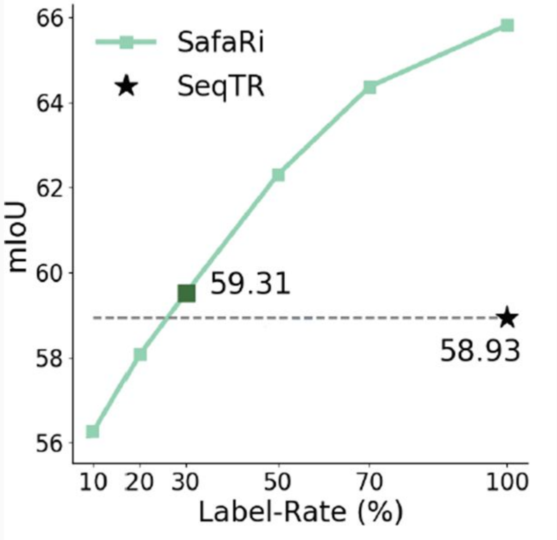
Impact of Retraining Steps



With increasing retraining steps (runs) in the bootstrapping pipeline, model becomes more confident which results in significant qualitative improvements in the predictions. SAFARI could not recognize the magazine partially behind laptop in the first step, it was successful in predicting the mask accurately in the final step - this shows the efficacy of the retraining stage.

Impact of Label Rates





- Quality of masks improve substantially with an increase in the ground-truth annotations.
- With just 30% annotations, SAFARI achieves 59.31 mIoU whereas fully-supervised SOTA (SeqTR) obtains 58.93 mIoU

Takeaways

1. Cross-modal fusion scheme is an **effective**, **flexible** and **parameter efficient** strategy and is particularly beneficial in the context of weakly-supervised RES task.

2. The AMCR component in X-FACT fosters **prediction of high-quality masks** by spatially constraining the cross-attention map within the target object boundary.

3. Bootstrapping pipeline enhances **self-labeling capabilities** in the system whereby filtered inferred masks are iteratively reutilized as pseudo-labels in the re-training phase, improving mIoU.

4. **Visual prompting** with red border together with the **spatial awareness component** in SPARC improves mIoU.

5. Mask Validity Filtering with SPARC **enhances the quality of training samples** in the retraining stage, which has a positive effect on the model performance.

Thank you!