Università degli Studi di Padova
Visual Intelligence Machine Perception Group

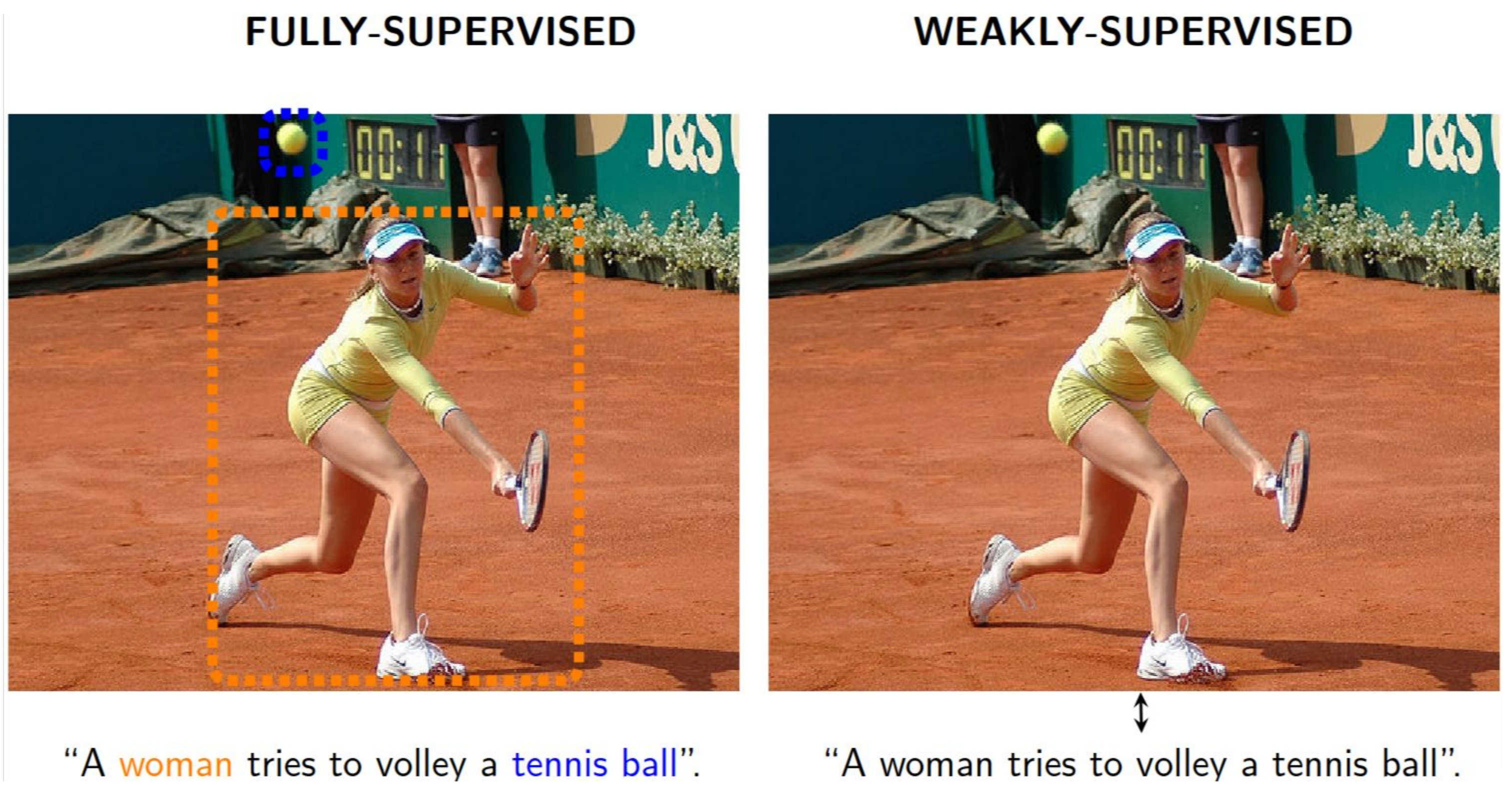
BMVC 2023

Weakly-Supervised Visual-Textual Grounding with Semantic Prior Refinement

D. Rigoni, L. Parolari, L. Serafini, A. Sperduti, L. Ballan
davide.rigoni@phd.unipd.it - luca.parolari@studenti.unipd.it

Paper

1. TASK AND PROBLEM

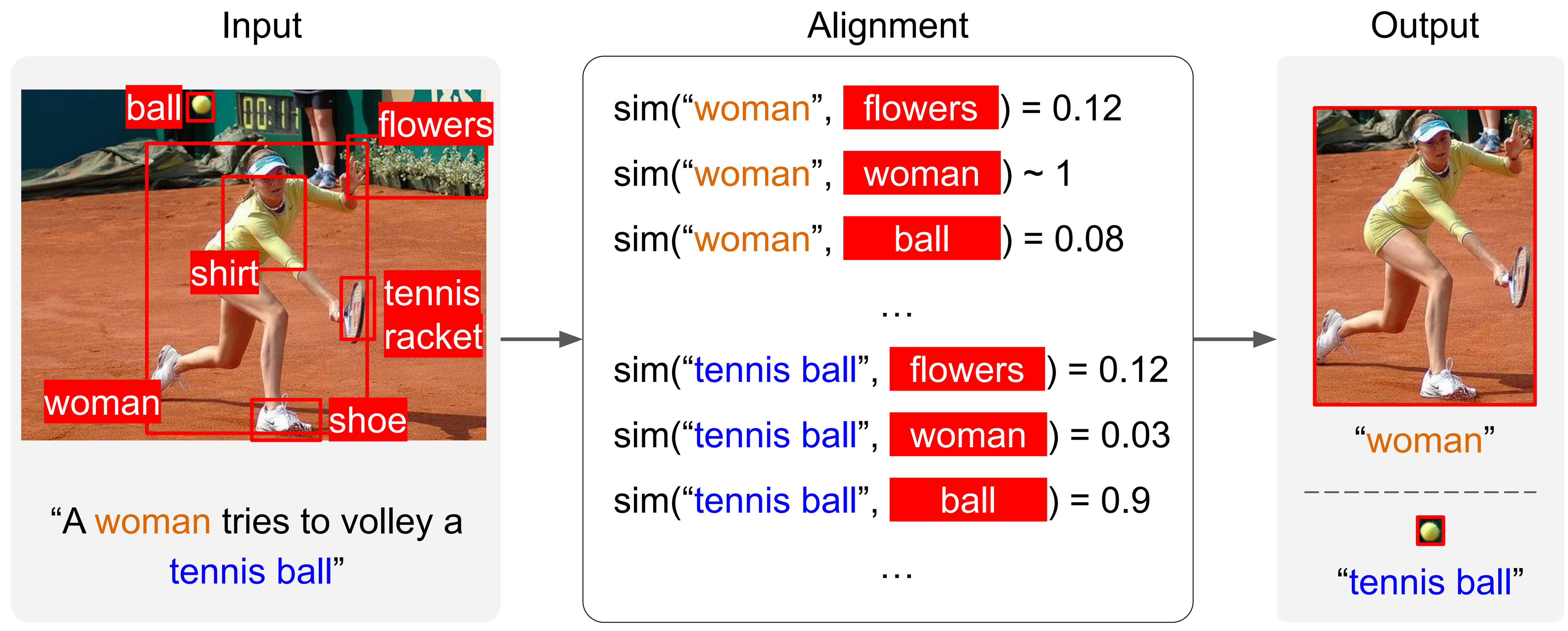


"A woman tries to volley a tennis ball". "A woman tries to volley a tennis ball".

Visual Grounding is the task of aligning the entity mentioned in a query with the respective portion of the image

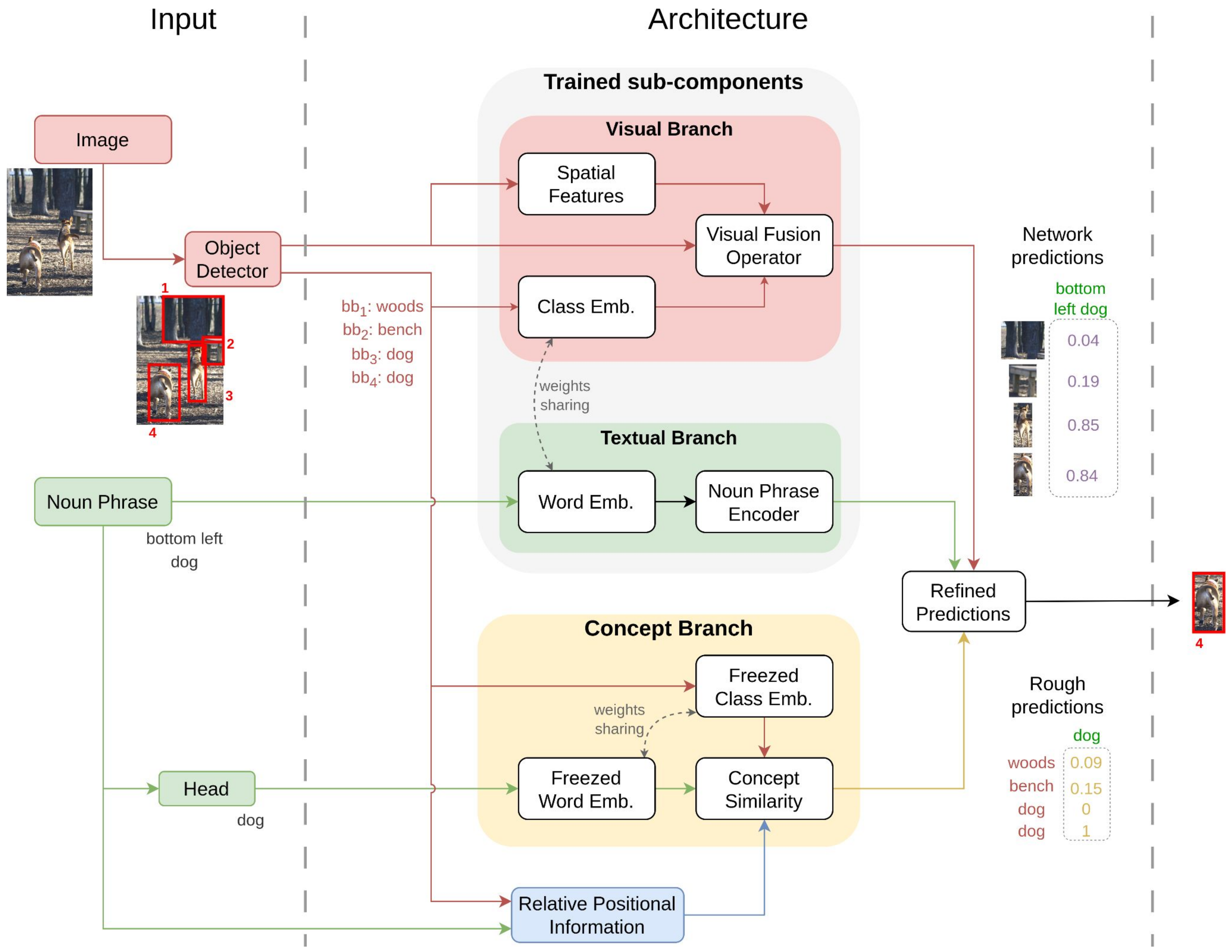
Issue: annotations are hard and expensive to collect

2. ALIGNING CONCEPTS



- In weakly-supervised setting, fine-grained annotations are not available at training time
- The object detector outputs the proposals and their categories
- Using word embedding we can grossly align phrases and proposals

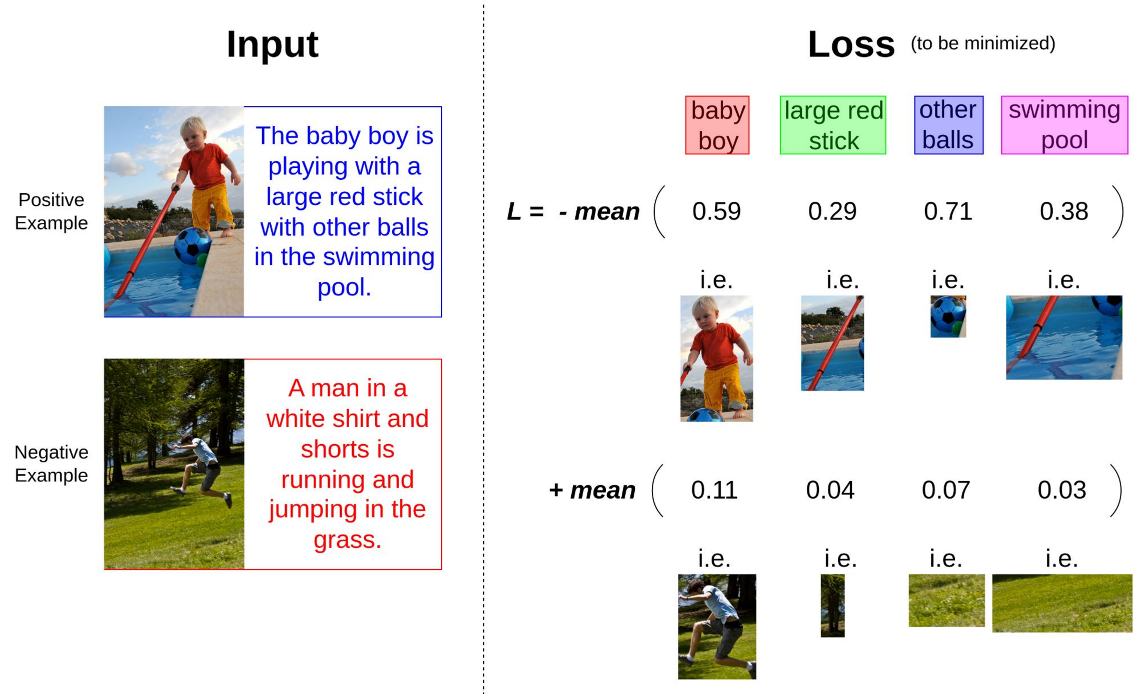
3. OUR APPROACH



Training

$$\mathcal{L} = - \underbrace{f_{pair}(\boldsymbol{I}, \mathrm{S})}_{\text{Positive example}} + \underbrace{f_{pair}(\boldsymbol{I'}, \mathrm{S})}_{\text{Negative example}}$$

- maximize the multimodal similarity f_{pair} of the image and its sentence (positive example)
- minimize f_{pair} between the same sentence and another image (negative example)



Negative example selection

The most similar example to the positive one, according to the sentence, in the minibatch. Therefore, the model focuses on details.

4. EXPERIMENTAL RESULTS

Model	Flickr30k E. (%) ↑ Acc.	Flickr30k E. (%) ↑ P. Acc.	ReferIt (%) ↑ Acc.	ReferIt (%) ↑ P. Acc.
Top-down Saliency	-	50.10	-	-
KAC Net	38.71	-	15.83	-
Semantic Self-Sup.	-	49.10	-	39.98
Anchored Transformer	33.10	-	13.61	-
Multi-level Multimodal	-	69.19	-	48.42
Align2Ground	-	71.00	-	-
Counterf. Resilience	48.66	-	-	-
MAF	61.4	-	-	-
Contrastive Learning	51.67	76.74	-	-
Grounding By Sep.	-	75.60	-	58.21
Relation-aware	59.27	78.60	37.68	58.96
Contrastive KL Distill.	53.10	-	38.39	-
EARN	38.73	-	36.86	-
RefCLIP	-	-	42.64	-
SimMaps	45.56	79.95	38.74	70.25
SPR baseline + CLIP (ours)	56.89	77.06	40.99	57.48
SPR model (ours)	**62.20**	**80.68**	**48.04**	**62.40**

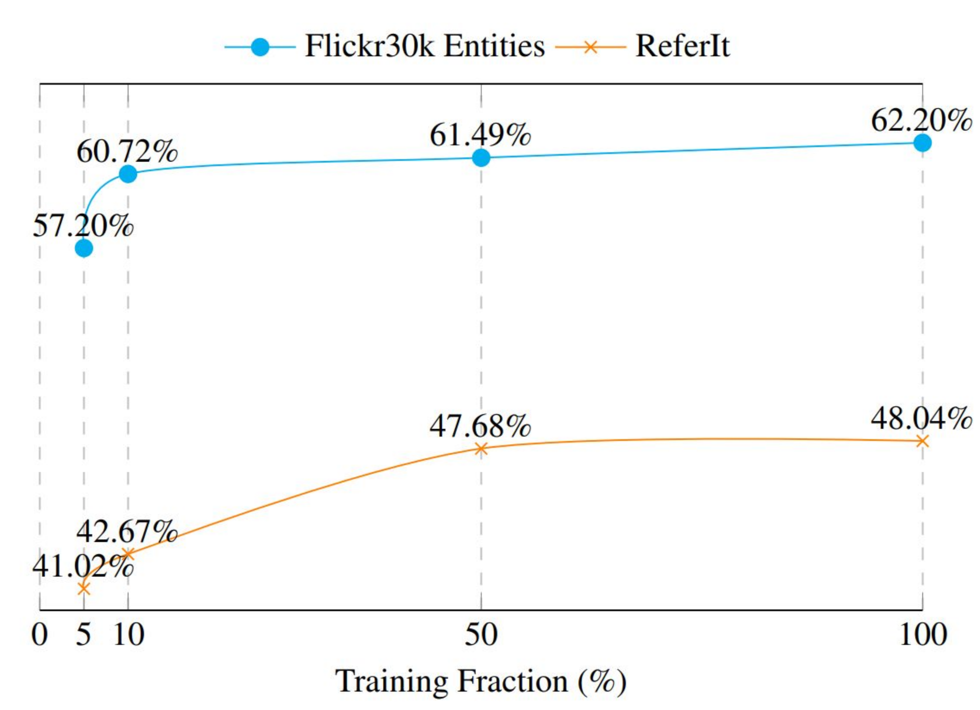
Results on Flickr30k Entities and ReferIt test sets. *Acc.* is the standard accuracy metric, while *P. Acc.* is the pointing game accuracy metric.

5. LOW-DATA SETTING

Accuracy results on Flickr30k Entities and ReferIt test set by our model trained in low-data environments.

The percentage refers to the fraction of the training set considered during training.

The model shows stable performances thanks to the concept branch.



6. MODEL ABLATION

Accuracy of our model's components. The Concept Branch contributes more to the final model performances.

Concept Branch	Trained Modules	Rel. Posit. Information	Flickr30k Entities (%)	ReferIt (%)
✖	✓	✖	23.52	15.03
✓	✖	✖	54.96	40.07
✓	✖	✓	55.02	42.69
✓	✓	✖	62.10	45.44
✓	✓	✓	**62.20**	**48.04**

7. CONCLUSION

1. We propose an untrained, zero-shot alignment module
2. Our model show comparable performance trained with 50% of data
3. Absolute improvement of 9.6% on ReferIt dataset