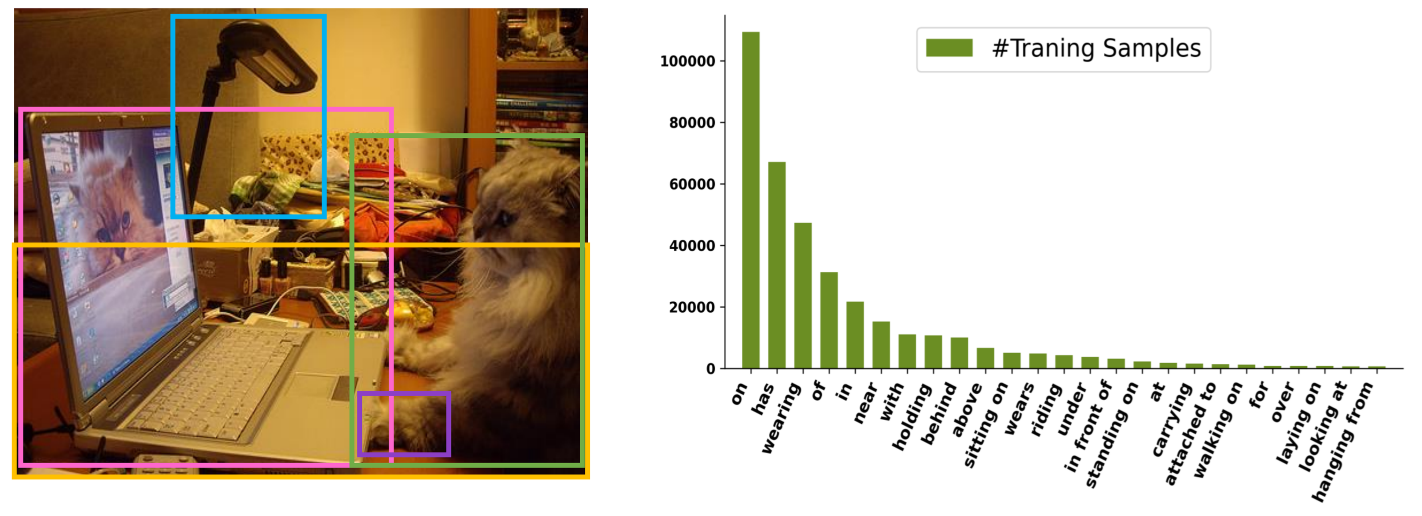
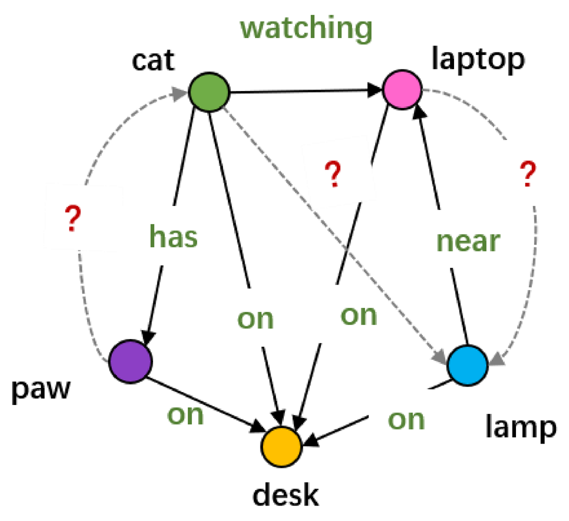
Rethinking the Evaluation of Unbiased Scene Graph Generation

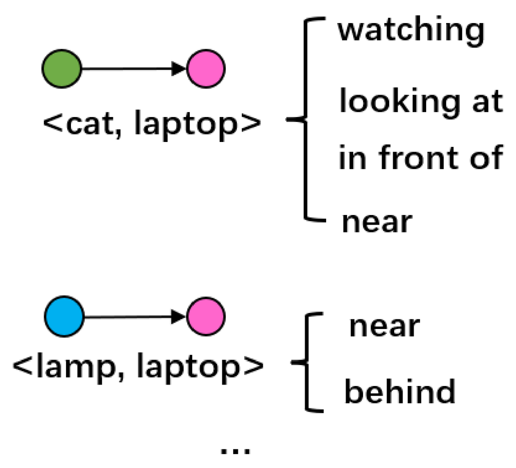
Xingchen Li[1], Long Chen[2*], Jian Shao[1], Shaoning Xiao[1], Songyang Zhang[3], Jun Xiao[1]
[1]Zhejiang University, [2]Columbia University, [3]University of Rochester

SGG Dataset Annotation



Long tailed distribution of predicates





Incompleteness **High Correlation**

Due to intrinsic complexity of the task and these inevitable annotation characteristics of SGG datasets, it is hard to properly evaluate SGG models.

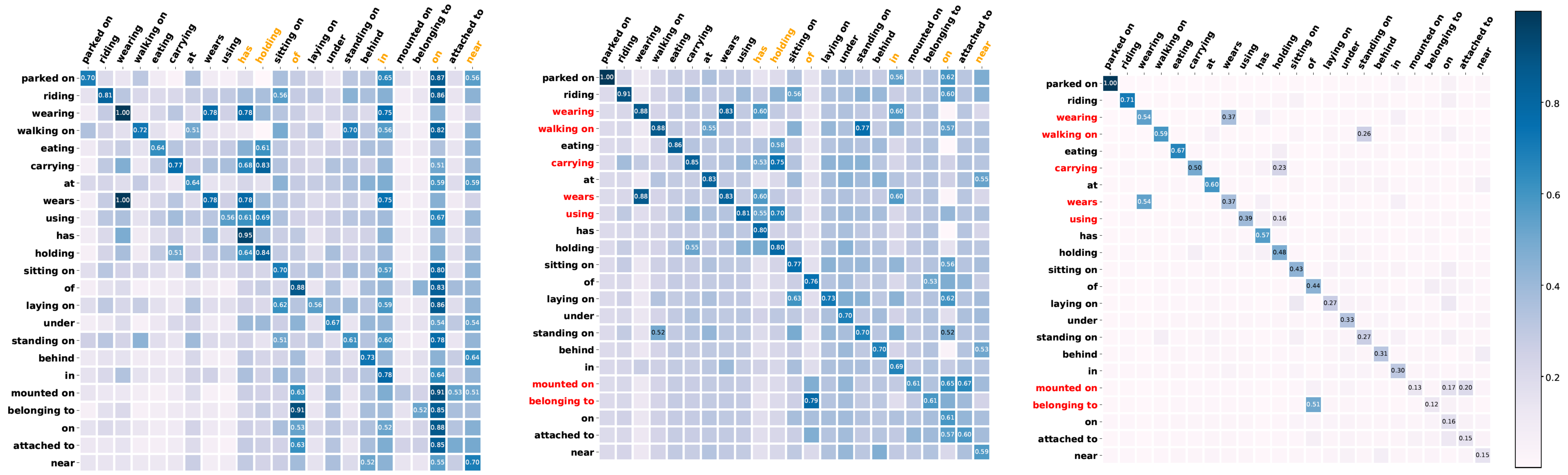
Predicate Knowledge on Objects

We investigate the intrinsic correlation between objects and predicates from a new perspective. We turn our attention to the distribution of object categories under each predicate category and obtain a new statistical prior, Predicate Knowledge on Objects (PKO). Directly aggregating PKO into the inference results of SGG models can improve unbiased performance of SGG models.

$$\mathbf{z}_{i,j} = \hat{\mathbf{z}}_{i,j} + \mathbf{b}_{i,j}, \qquad b_{i,j,k} = -\log\frac{\tilde{\mathbf{A}}^s_{k,i}}{\sum_{c\in C}\tilde{\mathbf{A}}^s_{c,i}} - \log\frac{\tilde{\mathbf{A}}^o_{k,j}}{\sum_{c\in C}\tilde{\mathbf{A}}^o_{c,j}}$$

Two Overlooked Issues and Suggestions

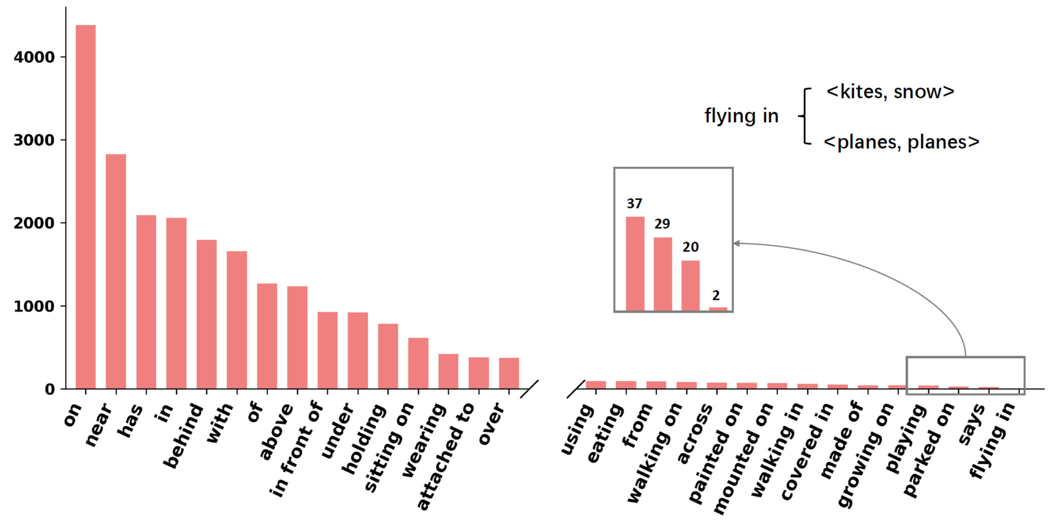
Ranking across categories may break *category Independence*.



Biased Output *Unbiased Output* *After Sofmax*

Assigning equal weights neglecting *compositional diversity*.

The compositional diversity of different predicates varies greatly. We devised a simple experiment and observe that *the predicates with limited compositional diversity have a stronger correlation with subject-object priors, which can be simply improved even without visual information.*



	#type$_{pair}$	mR@100	**Impro.%**
Motifs (N=0)	-	17.2	-
$N=1$ (flying in)	2	18.4	**1.2**
$N=2$ (says)	20	19.2	**2.0** +0.8
$N=3$ (parked on)	29	21.0	**3.8** +1.8
$N=4$ (playing)	37	20.9	**3.7** - 0.1
$N=5$ (growing on)	41	22.3	**5.1** +1.4
$N=6$ (made of)	42	22.9	**5.7** +0.6

Independent Mean Recall (IMR). We suggest independently ranking and output top-K (K = 10/20/50) predictions for each predicate category to calculate their own recall scores on this image.

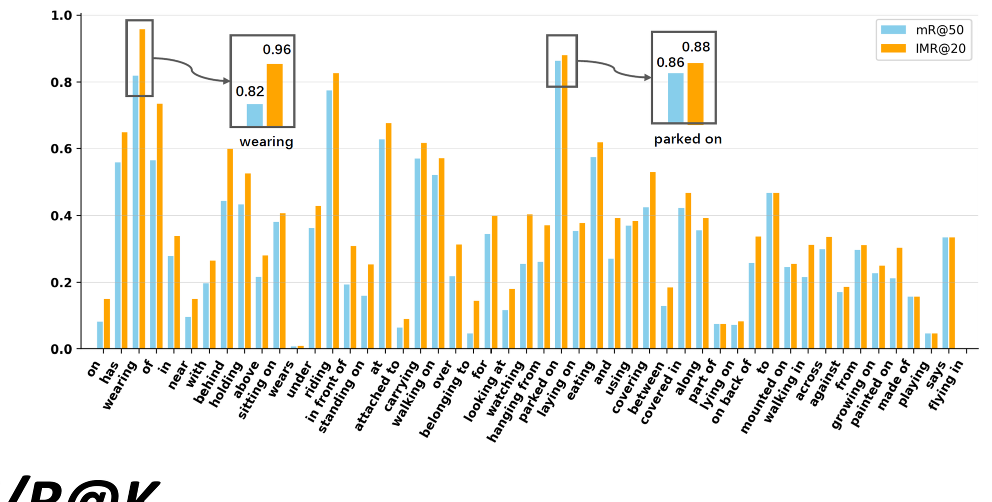
Weighted Independent Mean Recall (wIMR). We suggest reassigning weights to each predicate category c according to the complexity of their compositional space. We count the number of composed subject-object pairs n_c for each predicate category, and reassign weights to each predicate category c.

$$w(n_c) = \frac{n_c^{\tau}}{\sum_{k\in C} n_k^{\tau}}, \qquad wIMR = \sum_{c\in C} w(n_c) \times IMR(c)$$

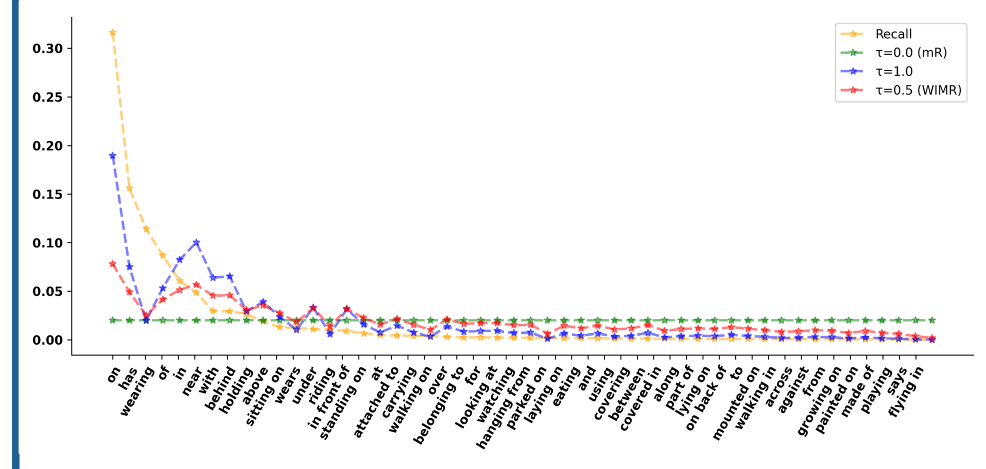
Experimental Results

IMR@K vs. mR@K.

Compared to mR@K, IMR@K provides a more fair score for those over-estimated predicate categories.



wIMR@K vs. mR@K/R@K..



By considering both data distribution and compositional diversity, wIMR manages to assign high weights to predicates with rich semantics and low weights to predicates with simple semantics。

Benchmarking SOTA Unbiased SGG Methods and PKO

	Method	PredCls R@20 / 50 / 100	PredCls mR@20 / 50 / 100	PredCls IMR@10 / 20 / 50	PredCls wIMR@10 / 20 / 50
	PKO	10.10 / 15.51 / 19.44	17.86 / 27.38 / 33.93	**38.03** / **39.02** / **40.90**	29.75 / 30.79 / 31.38
Motifs	Baseline	**59.19** / **65.59** / **67.30**	12.62 / 15.96 / 17.23	14.53 / 16.11 / 17.45	24.80 / 27.99 / 30.97
	TDE	33.35 / 45.89 / 51.24	17.85 / 24.77 / 28.72	27.33 / 29.52 / 31.06	32.64 / 35.59 / 37.72
	DLFE	44.28 / 50.33 / 51.99	21.86 / 26.81 / 28.53	26.37 / 27.92 / 29.03	31.49 / 34.02 / 35.73
	NICE	48.15 / 55.14 / 57.15	23.67 / 29.83 / 32.24	28.35 / 31.31 / 33.12	31.61 / 35.26 / 37.84
	Reweight	26.57 / 36.08 / 40.39	23.98 / 30.79 / 34.48	**35.39** / **36.58** / **36.98**	**36.13** / **37.52** / 38.05
	RTPB(CB)	36.58 / 42.64 / 44.36	**27.61** / **32.78** / **34.57**	30.55 / 33.30 / 35.02	33.63 / 37.16 / **39.32**
	PKO	49.08 / 55.95 / 58.18	24.98 / 31.44 / 33.98	29.73 / 32.46 / 34.21	31.77 / 35.51 / 38.51
VCTree	Baseline	**59.72** / **65.86** / **67.50**	13.26 / 16.82 / 18.12	15.36 / 16.99 / 18.30	25.31 / 28.58 / 31.58
	TDE	34.48 / 44.89 / 49.20	19.07 / 25.61 / 29.13	27.24 / 29.46 / 31.03	32.59 / 35.70 / 38.05
	DLFE	45.35 / 51.21 / 52.75	22.53 / 27.36 / 28.86	26.46 / 28.28 / 29.30	31.49 / 34.27 / 36.01
	NICE	48.38 / 55.03 / 56.92	24.42 / 30.74 / 33.01	29.03 / 32.01 / 33.93	31.72 / 35.41 / 38.12
	Reweight	28.66 / 35.62 / 37.90	**28.64** / **34.93** / **37.28**	**34.70** / **36.95** / **38.30**	**34.41** / 36.94 / 38.42
	RTPB(CB)	36.65 / 42.39 / 43.95	28.64 / 33.41 / 35.11	30.57 / 33.47 / 33.50	33.22 / **36.99** / **39.50**
	PKO	49.39 / 56.06 / 58.18	26.06 / 32.20 / 34.61	30.61 / 33.41 / 35.29	31.98 / 35.73 / 38.88

	Method	SGCls R@20 / 50 / 100	SGCls mR@20 / 50 / 100	SGCls IMR@10 / 20 / 50	SGCls wIMR@10 / 20 / 50
	PKO	7.26 / 10.48 / 12.46	11.23 / 16.77 / 20.13	20.51 / 22.47 / 23.11	16.04 / 17.13 / 17.68
Motifs	Baseline	**36.39** / **39.59** / **40.35**	7.44 / 9.09 / 9.63	7.69 / 8.65 / 9.57	13.65 / 15.70 / 17.76
	TDE	20.46 / 26.31 / 28.78	9.78 / 13.21 / 15.07	13.98 / 15.22 / 16.29	16.95 / 18.69 / 20.16
	DLFE	26.63 / 29.79 / 30.61	13.23 / 15.66 / 16.48	14.75 / 15.82 / 16.57	17.96 / 19.65 / 20.85
	NICE	29.48 / 33.06 / 34.05	13.63 / 16.67 / 17.88	14.82 / 16.76 / 18.02	17.00 / 19.35 / 21.07
	Reweight	18.67 / 23.49 / 25.47	13.49 / 16.75 / 18.34	**17.82** / **18.69** / **19.23**	**19.13** / 20.38 / 21.14
	RTPB(CB)	22.66 / 25.85 / 26.65	**15.77** / **18.16** / 18.97	16.08 / 17.68 / 18.81	18.51 / **20.61** / **21.99**
	PKO	30.41 / 33.99 / 35.05	14.06 / 17.59 / **19.12**	15.87 / 17.49 / 18.66	17.43 / 19.72 / 21.72
VCTree	Baseline	**42.09** / **45.80** / **46.73**	9.09 / 11.28 / 12.04	9.72 / 10.89 / 11.94	16.52 / 18.90 / 21.26
	TDE	23.48 / 31.17 / 34.59	10.36 / 14.47 / 16.72	15.81 / 17.32 / 18.48	19.71 / 21.73 / 23.35
	DLFE	30.09 / 33.85 / 34.80	16.17 / 19.23 / 20.20	18.81 / 20.04 / 20.62	21.75 / 23.58 / 24.65
	NICE	33.77 / 37.84 / 38.99	16.14 / 20.03 / 21.29	17.89 / 19.91 / 21.59	20.16 / 22.77 / 24.80
	Reweight	20.44 / 24.66 / 25.97	**18.72** / **22.88** / **24.19**	**22.14** / **23.66** / **24.58**	22.04 / 23.93 / 25.07
	RTPB(CB)	26.06 / 29.63 / 30.61	18.67 / 21.41 / 22.52	19.64 / 21.41 / 22.63	**22.04** / **24.41** / **25.93**
	PKO	35.01 / 39.10 / 40.40	18.43 / 22.27 / 23.74	19.88 / 21.72 / 23.01	21.14 / 23.77 / **26.09**