We thank the reviewers for the time and attention invested in these reviews. We address the reviewers' remarks below.

> Include discussion of [A,B,C,D] and empirical comparison to [A,B,C, 21]

We will add empirical comparison with [A,B,C] (the authors of [21] have not released their model, nor did they present evaluation on the MSCOCO dataset). The following are the conceptual comparisons to be added to Related Work: [A] does not consider the size and position of detected regions in their algorithm. [B] classifies the spatial relationships of two boxes into 11 classes, such as "inside", "cover" or "overlap". Our approach directly utilizes the size ratio and difference of x- and y-location to compute the box relationship, which implicitly encodes the mentioned properties. [C] does incorporate object spatial relationships, but does not use visual geometry to do so. [D] does successfully learn relationships between all object pairs of objects; however, it is not comparable to the image captioning task as it does not provide a single caption describing the overall image scene. In addition, all of these works are using LSTMs while our algorithm uses the transformer architecture for the caption generation.

> Include evaluation on the online MSCOCO test server.

The MSCOCO test server results from our model are shown below. Note that the model used in this submission was trained on the Karpathy train split, rather than the standard practice of training it on Train+Val. For a fairer comparison, we will do so in the camera-ready paper version.

test set	CIDEr-D	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L
c5 / c40	123.6 / 125.4	80.2 / 94.1	64.7 / 88.3	50.0 / 79.3	37.8 / 68.7	28.5 / 37.4	58.1 / 72.9

> If time permits, include a human evaluation, comparing the proposed model to strong baselines or SoTA models.

A comparison based on human judgments should give additional insights, and we'll add it to the camera-ready paper.

> Discuss failure modes of the proposed method.

We will analyze and discuss failure modes along with illustrative examples, space permitting. We manually reviewed results from our best model over 100 randomly sampled images from the MSCOCO test set (and will extend the analysis to the complete test set for the camera-ready submission). Out of the 62 observed errors, 58% pertained to objects or things (24% missed, 24% wrong, 10% extraneous), 21% to relations (5% missed, 16% wrong), 16% to attributes (5% missed, 8% wrong counts, 1.5% wrong colors, 1.5% extraneous), and 5% to syntactic errors.

> Do all methods in Table 1 use the same beam size? If not, what fraction of the improvement is attributed to performing beam search? The authors should compare the methods with same beam size, and clarify it on Tab. 1

Our best results were obtained with beam size 2, in consistency with other research on Image Captioning optimization [20] (appendix A). Only in Tab. 1, for a fair comparison with other models in the literature, we present our result with the same beam size of 5 that they used to communicate their performance, indicating it in the table's caption. We will make this distinction more clear in the Implementation Details Section, by specifying the setting for each experiment. Finally, the last line of our ablation study in Tab. 3 addresses the fraction of the improvement attributed to beam search.

> There is a difference in the scores for the Up-Down baseline between Tab. 3 and Tab. 1. Is it because self-critical training was not used for Tab. 3? Are the splits different?

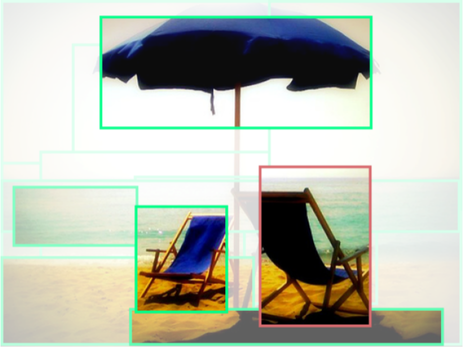
That is correct, self-critical training was not used in Tab. 3. In addition, we used our own implementation of the Up-Down algorithm for the baseline results in Tab. 3. The splits are the same in all the tables.

> For the proposed features, the improvement on "size" sub-score of SPICE is not statistically significant. In such a scenario, it would be help to study different relative location features for the two bounding boxes.

We experimented with different object geometric features, as well as different ways of fusing these with the object appearance features. We did not try the suggested technique from Interaction Pattern of Chao et al. in "Learning to detect human-object interactions", which does encode the relative size of pairs of objects. Since this method scales the bounding box 0-1 masks of pairs of objects, it may not address the issue identified by the SPICE absolute object size metric. But we are enthusiastic to check if we can get a boost following the suggested technique.

> A visualization for the attention weights for an appropriate layer (say for the qualitative examples in Tab. 7) may be useful to demonstrate that the network does indeed learn spatial relationship information.

We generated the following visualization of the self-attention in our proposed Object Relation Transformer, here displayed for one of the relation example images (averaged across attention heads, and self-attention layers). The detected object transparency is proportional to the attention weight with respect to the chair outlined in red. This image will be discussed in the paper.



Generated Caption: *two beach chairs under an umbrella on the beach*

> What happens on other vision and language tasks?

We are considering applying this approach to other tasks (HOI and VQA).