A Appendix

A.1 Limitations and Future Works

The empirical examination of our framework reveals the following limitations:

Pretrained representations. Like prior arts, our approach relies on pretrained object detector and a language model to represent regions and caption-words. Ideally, we would expect to learn from scratch or improve existing region and word representations directly from image-caption data.

Need for fully-labeled validation set. In Fig. 4, we observe that an early stopping based on the validation performance is required to choose the best model for phrase grounding. While this is common practice for weakly supervised learning [10] and the Flickr30K Entities validation set we use is 80× smaller than the COCO training set, this translates to using full supervision for a small set of images.

Bounds on MI. While $\log(K) - \mathcal{L}_{\texttt{img}}$ in Eq. 8 is a valid lower bound on MI, our $\log(K) - \mathcal{L}_{\texttt{lang}}$ in Eq. 9 is no longer a lower bound on MI as it oversamples negative words related to a caption. A valid bound would involve random sampling of captions from the training data however our context-preserving negative captions lead to much better performance.

A.2 Advantages of Context-Preserving Negative Sampling

Commonly used strategies for negative sampling for contrastive learning include randomly sampling captions from the training data as negatives or mining hard-negatives from a randomly sampled mini-batch. In our experiments (Tab. 2), random sampling showed no significant gains over a model trained without negative captions. This is because the sampled negatives often have an entirely different context as compared to the image and the positive caption which makes it too easy for the model to produce a low compatibility score for these negatives.

In contrast, contrast-preserving negative sampling shows significant gains over random sampling (76.74% *vs.* 66.89% pointing accuracy). This is because we construct harder negative captions which yield a more informative training signal than random sampling. We construct negatives by substituting only a single word in the caption while preserving the context from the positive caption. The substitutions are further chosen to be plausible given the context while discarding likely synonyms and hypernyms. Unlike random sampling approaches whose success depends on the occurrence of informative negative captions in the training data and the likelihood of sampling such negatives for a positive caption in the same minibatch, our approach can construct effective negatives for any positive caption.

A.3 Relation between our query-key-value attention and self-attention in Transformers

Our query-key-value attention mechanism is related to the attention mechanism used in transformer-based [41] architectures like BERT [12]. Transformers use the mechanism for self-attention where queries, keys, and values are computed for each word in the input sentence and the attention scores are used for contextualization. In contrast, we use the attention mechanism for word-region alignment. Specifically, we compute queries for each contextualized word, keys for each region, and values for regions as well as words (using separate value networks for regions and words).

A.4 Comparison to Align2Ground

While we use the same visual features as the previous SOTA, Align2Ground [11], the two approaches use different textual features. While Align2Ground uses a bi-GRU, we chose BERT, a transformer-based language model which became more prevalent (as opposed to RNN-based) in the vision-language community. To estimate the gain due to pretrained language representations, Tab. 2 compares the grounding performance of randomly initialized BERT (57.37%) to that of pretrained BERT (66.89%). Negative sampling brings further gains (76.74%).