# Learning to Generate Grounded Visual Captions without Localization Supervision (Supplementary Material)

Chih-Yao Ma[1,3], Yannis Kalantidis[⋆2], Ghassan AlRegib[1], Peter Vajda[3], Marcus Rohrbach[3], Zsolt Kira[1]
[1]Georgia Tech, [2]NAVER LABS Europe, [3]Facebook

{cyma,alregib,zkira}@gatech.edu, yannis.kalantidis@naverlabs.com, {cyma,vajdap,mrf}@fb.com

## Appendix

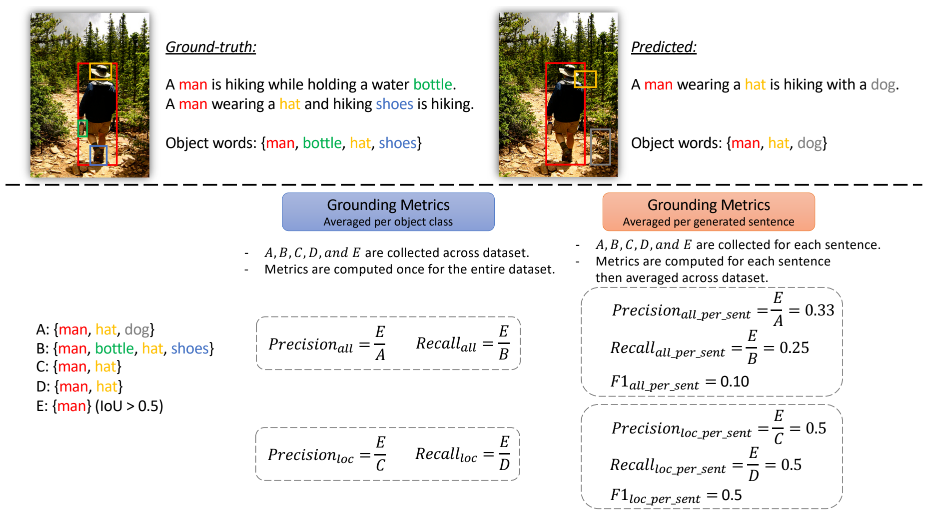


Fig. 1: Illustration of Grounding metrics.

## A Grounding Evaluation Metrics illustrated

To help better understand the grounding evaluation metrics used in this work, we illustrated the grounding evaluation metrics in Figure 1.

We define the number of object words in the generated sentences as A, the number of object words in the GT sentences as B, the number of correctly predicted object words in the generated sentences as C and the counterpart in the

⋆ Work done while at Facebook.

| Method | Captioning Evaluation | | | | | Grounding Evaluation | | | |
|---|---|---|---|---|---|---|---|---|---|
| | B@1 | B@4 | M | C | S | $F1_{all}$ | $F1_{loc}$ | $F1_{all\_per\_sent}$ | $F1_{loc\_per\_sent}$ |
| Baseline | 69.1 | 26.0 | 22.1 | 59.6 | 16.3 | 4.08 | 11.83 | 13.20 | 31.83 |
| Cyclical | 69.4 | 27.4 | 22.3 | 61.4 | **16.6** | 4.98 | 13.53 | 15.03 | 35.54 |
| Cyclical (zero-loss) | 69.7 | 27.0 | 22.2 | 60.1 | 16.5 | **5.14** | **14.32** | 15.36 | 36.33 |
| Cyclical (zero-representation) | **69.9** | **27.5** | **22.4** | **62.0** | **16.6** | 5.13 | 13.99 | **16.30** | **38.45** |

Table 1: Performance comparison on the Flickr30k Entities **test set**. All results are averaged **across five runs**.

GT sentences as D, and the number of correctly predicted and localized words as E. A region prediction is considered correct if the object word is correctly predicted and also correctly localized (*i.e.,* IoU with GT box > 0.5). We then compute two version of the precision and recall as $\text{Prec}_{\text{all}} = \frac{E}{A}$, $\text{Rec}_{\text{all}} = \frac{E}{B}$, $\text{Prec}_{\text{loc}} = \frac{E}{C}$, and $\text{Rec}_{\text{loc}} = \frac{E}{D}$.

The original grounding evaluation metric proposed in GVD [4] average the grounding for each object class. We additionally calculate the grounding accuracy for each generated sentence as demonstrated in the figure. From this example, we can see that while $Precision_{all}$ counts *dog* as a wrong prediction for the *dog* object class, the $Precision_{loc}$ only cares if *man* and *hat* are predicted and correctly localizer (IoU > 0.5).

# B Additional Quantitative Analysis

**Can words that are not visually-groundable be handled differently?** In the proposed method, all the words are handled the same regardless of whether they are visually-groundable or not, *i.e.,* the localizer is required to use *all* generated words at each step in a sentence to localize regions in the image. Yet, typically words that are nouns or verbs are more likely to be grounded, and words like "a", "the", *etc*, are not visually-groundable.

We explored a few method variants to handle nouns and verbs differently. Mainly, we explored with two variants.

- **Cyclical (zero-loss):** the reconstruction loss is only computed when the target word is either a noun or a verb.
- **Cyclical (zero-representation):** the localized region representation will be invalid (set to zero) if the target word is neither nouns nor verbs.

The experimental results are shown in Table 1, 2, and 3. For the first variant, Cyclical (zero-loss), we observed that the captioning performance stays the same while grounding accuracy has a small improvement. On the other hand, for the second variant, Cyclical (zero-representation), we can see that all captioning scores are improved over baseline with CIDEr improved 2.4 (see Table 1). We can also see that grounding accuracy on per sentence basis further improved

| Method | B@1 | B@4 | M | C | S | $F1_{all}$ | $F1_{loc}$ | $F1_{all\_per\_sent}$ | $F1_{loc\_per\_sent}$ |
|---|---|---|---|---|---|---|---|---|---|
| | Captioning Evaluation | | | | | Grounding Evaluation | | | |
| Baseline | 23.2 | 2.22 | 10.8 | 45.9 | **15.1** | 3.75 | 12.00 | 9.41 | 31.68 |
| Cyclical | 23.7 | 2.45 | 11.1 | 46.4 | 14.8 | **4.68** | **15.84** | **12.60** | **44.04** |
| Cyclical (zero-representation) | **23.9** | **2.58** | **11.2** | **46.6** | 14.8 | 4.48 | 15.01 | 11.53 | 40.30 |

Table 2: Performance comparison on the ActivityNet-Entities **val set**. All results are averaged **across five runs**.

| Method | B@1 | B@4 | M | C | S | $F1_{all}$ | $F1_{loc}$ | $F1_{all\_per\_sent}$ | $F1_{loc\_per\_sent}$ |
|---|---|---|---|---|---|---|---|---|---|
| | Captioning Evaluation | | | | | Grounding Evaluation | | | |
| **Unrealistically perfect object detector** | | | | | | | | | |
| Baseline | 75.1 | 32.1 | 25.2 | 76.3 | 22.0 | 20.82 | 48.74 | 43.21 | 77.81 |
| Cyclical | **76.7** | **32.8** | **25.8** | **80.2** | **22.7** | 25.27 | 54.54 | 46.98 | 81.56 |
| Cyclical (zero-representation) | 75.8 | 32.2 | 25.6 | 79.0 | 22.4 | **25.65** | **55.81** | **48.99** | **85.99** |

Table 3: Grounding performance when using better object detector on the Flickr30k Entities **test** set (results are averaged three runs).

as well. We then conducted further experiments on both ActivityNet-Entities and Flickr30k Entities with *unrealistically perfect object detector* (see Table 2 and 3), but the improvements however are not consistent. In summary: on the Flickr30k Entities test set, we observed that CIDEr is better and grounding per sentence better, on the ActivityNet-Entities val set, the captioning performances are about the same but grounding accuracy became worse, and on the Flickr30k Entities test set with unrealistically perfect object detector, captioning performances are slightly worse but grounding accuracy improved. We thus keep the most general variant "Cyclical" which treats all words equally.

**Will a non-linear localizer performs better?** In practice, our localizer is a single fully-connected layer. It is possible to replace it with a non-linear layer, *e.g.,* multi-layer perceptron (MLP). We however observed that both captioning and grounding accuracy reduced if a MLP is used as the localizer (see Table 4).

**Weighting between decoding and reconstruction losses.** The weighting between the two losses was chosen with a grid search on the val set. We report the experimental results on Flickr30k Entities val set in Table 5. We can see that when comparing to the baseline, all different loss weightings consistently improved both captioning and grounding accuracy. Unless further specified, we use default (0.5, 0.5) weighting for the two losses, except (0.6, 0.4) for the final result on Flickr30k Entities test set.

## C Additional Qualitative Results

In Figure 3, 4, 5, 6, 7, 8, 9, and 10, we illustrated the sequence of attended image region when generating each word for a complete image description. At each step, only the top-1 attended image region is shown. This is the same as

| Method | Captioning Evaluation B@1 | B@4 | M | C | S | Grounding Evaluation $F1_{all}$ | $F1_{loc}$ | $F1_{all\_per\_sent}$ | $F1_{loc\_per\_sent}$ |
|---|---|---|---|---|---|---|---|---|---|
| Cyclical | **69.4** | **27.4** | **22.3** | **61.4** | **16.6** | **4.98** | **13.53** | **15.03** | **35.54** |
| Cyclical (MLP Localizer) | 69.2 | 26.4 | 22.0 | 58.7 | 16.2 | 4.40 | 12.77 | 13.97 | 33.40 |

Table 4: Performance comparison on the Flickr30k Entities **test set** using FC or MLP as the localizer. All results are averaged **across five runs**.

| $(\lambda_1, \lambda_2)$ | Captioning Evaluation B@1 | B@4 | M | C | S | Grounding Evaluation $F1_{all}$ | $F1_{loc}$ | $F1_{all\_per\_sent}$ | $F1_{loc\_per\_sent}$ |
|---|---|---|---|---|---|---|---|---|---|
| baseline | 69.7 | 26.7 | 22.3 | 61.1 | 16.1 | 4.61 | 13.11 | 12.41 | 30.61 |
| (0.8, 0.2) | 70.3 | 27.9 | 22.4 | 62.2 | 16.5 | 4.96 | 13.95 | 13.95 | 33.49 |
| (0.6, 0.4) | 70.4 | 28.0 | 22.4 | 62.7 | 16.3 | 5.04 | 13.92 | 14.46 | 34.95 |
| (0.5, 0.5) | 70.2 | 27.9 | 22.5 | 62.3 | 16.5 | 4.93 | 13.70 | 14.28 | 34.62 |
| (0.4, 0.6) | 69.8 | 27.6 | 22.5 | 62.3 | 16.4 | 4.97 | 13.67 | 14.97 | 36.31 |
| (0.2, 0.8) | 69.5 | 27.7 | 22.3 | 61.4 | 16.1 | 5.07 | 14.05 | 15.41 | 37.63 |

Table 5: Performance comparison on the Flickr30k Entities **val set** with different weightings on decoding and reconstruction losses. All results are averaged **across five runs**.

how the grounding accuracy is measured. Please see the description for Figure 3 - 10 for further discussions on the qualitative results.

## D Additional Implementation Details

**Region proposal features.** We use a Faster-RCNN model [3] pre-trained on Visual Genome [2] for region proposal and feature extraction. In practice, besides the region proposal features, we also use the Conv features (*conv4*) extracted from an ImageNet pre-trained ResNet-101. Following GVD [4], the region proposals are represented using the *grounding-aware region encoding*, which is the concatenation of i) region feature, ii) region-class similarity matrix, and iii) location embedding.

For region-class similarity matrix, we define a set of object classifiers as $\boldsymbol{W}_c$, and the region-class similarity matrix can be computed as $M_s = \text{softmax}(W_c^\top \boldsymbol{R})$, which captures the similarity between regions and object classes. We omit the ReLU and Dropout layer after the linear embedding layer for clarity. We initialize $\boldsymbol{W}_c$ using the weight from the last linear layer of an object classifiers pre-trained on the Visual Genome dataset [2].

For location embedding, we use 4 values for the normalized spatial location. The 4-D feature is then projected to a $d_s = 300$-D location embedding for all the regions.

**Software and hardware configuration.** Our code is implemented in PyTorch. All experiments were ran on the 1080Ti, 2080Ti, and Titan Xp GPUs.

| Method | Grounding supervision | B@1 | B@4 | M | C | S | $F1_{all}$ | $F1_{loc}$ | $F1_{all\_per\_sent}$ | $F1_{loc\_per\_sent}$ |
|---|---|---|---|---|---|---|---|---|---|---|
| | | Captioning Evaluation | | | | | Grounding Evaluation | | | |
| **Unrealistically perfect object detector** | | | | | | | | | | |
| Baseline | ✓ | 75.6 | 32.0 | 25.3 | 75.6 | 22.3 | 23.19 (+100%) | 52.83 (+100%) | 51.43 (+100%) | 90.76 (+100%) |
| Baseline | | 75.1 | 32.1 | 25.2 | 76.3 | 22.0 | 20.82 (+0%) | 48.74 (+0%) | 43.21 (+0%) | 77.81 (+0%) |
| Cyclical | | **76.7** | **32.8** | **25.8** | **80.2** | **22.7** | **25.27** (+188%) | **54.54** (+142%) | **46.98** (+46%) | **81.56** (+29%) |
| **Grounding-biased object detector** | | | | | | | | | | |
| Baseline | ✓ | 65.9 | 23.4 | 21.3 | 53.3 | 15.5 | 8.23 (+100%) | 23.95 (+100%) | 28.06 (+100%) | 66.96 (+100%) |
| Baseline | | **66.1** | **23.5** | **21.2** | **52.4** | **15.4** | 5.95 (+0%) | 17.51 (+0%) | 18.11 (+0%) | 42.84 (+0%) |
| Cyclical | | 65.5 | 23.3 | **21.2** | 52.0 | **15.4** | **6.87** (+40%) | **19.65** (+33%) | **20.82** (+27%) | **50.25** (+31%) |

Table 6: Grounding performance when using better object detector on the Flickr30k Entities **test** set (results are averaged three runs). Fully-supervised method is used as upper bound, thus its numbers are not bolded.

| Method | Grounding supervision | B@1 | B@4 | M | C | S | $F1_{all}$ | $F1_{loc}$ |
|---|---|---|---|---|---|---|---|---|
| | | Captioning Evaluation | | | | | Grounding Evaluation | |
| Masked Transformer [5] | | 22.9 | 2.41 | 10.6 | 46.1 | 13.7 | - | - |
| Bi-LSTM+TempoAttn [5] | | 22.8 | 2.17 | 10.2 | 42.2 | 11.8 | - | - |
| GVD (w/o SelfAttn) [4] | | 23.1 | 2.16 | **10.8** | 44.9 | 14.9 | 3.73 | 11.7 |
| GVD [4] | ✓ | 23.6 | 2.35 | 11.0 | 45.5 | 14.7 | 7.59 | 25.0 |
| Baseline | ✓ | 23.1 | 2.28 | 10.8 | 45.6 | 14.7 | 7.66 (+100%) | 25.7 (+100%) |
| Baseline | | 23.2 | 2.17 | **10.8** | 46.2 | **15.0** | 3.60 (+0%) | 12.3 (+0%) |
| Cyclical | | **23.4** | **2.43** | **10.8** | **46.6** | 14.3 | 4.70 (+27%) | 15.6 (+29%) |

Table 7: Performance comparison on the ActivityNet-Entities **test set**. Grounding evaluation metrics on per generated sentences are not available on the test server.

**Network architecture.** The embedding dimension for encoding the sentences is 512. We use a dropout layer with ratio 0.5 after the embedding layer. The hidden state size of the Attention and Language LSTM are 1024. The dimension of other learnable matrices are: $\boldsymbol{W}_e \in \mathbb{R}^{d_v \times 512}$, $\boldsymbol{W}_a \in \mathbb{R}^{1024 \times 512}$, $\boldsymbol{W}_{aa} \in \mathbb{R}^{512 \times 1}$, $\boldsymbol{W}_o \in \mathbb{R}^{1024 \times d_v}$, $\boldsymbol{W}_l \in \mathbb{R}^{512 \times 512}$, where the vocabulary size $d_v$ is 8639 for Flickr30k Entities and 4905 for ActivityNet-Entities.
**Training details.** We train the model with ADAM optimizer [1]. The initial learning rate is set to $1e-4$. Learning rates automatically drop by 10x when the CIDEr score is saturated. The batch size is 32 for Flickr30k Entities and 96 for ActivityNet-Entities. We learn the word embedding layer from scratch for fair comparisons with existing work [4]. The hyper-parameters $\lambda_1$ and $\lambda_2$ are set to 0.5 after hyper-parameter search between 0 and 1.
**Flickr30k Entities.** Images are randomly cropped to $512 \times 512$ during training, and resized to $512 \times 512$ during inference. Before entering the proposed cyclical training regimen, the decoder was pre-trained for about 35 epochs. The total training epoch with the cyclical training regimen is around 80 epochs. The total training time takes about 1 day.
**ActivityNet-Entities.** Before entering the proposed cyclical training regimen, the decoder was pre-trained for about 50 epochs. The total training epoch with

| Method | Captioning Evaluation B@1 | B@4 | M | C | S | Grounding Evaluation $F1_{all}$ | $F1_{loc}$ | $F1_{all\_per\_sent}$ | $F1_{loc\_per\_sent}$ |
|---|---|---|---|---|---|---|---|---|---|
| Baseline | 69.1±0.6 | 26.0±0.6 | 22.1±0.3 | 59.6±0.6 | 16.3±0.2 | 4.08±0.40 | 11.83±1.27 | 13.20±0.60 | 31.83±1.36 |
| Cyclical | **69.4±0.4** | **27.4±0.1** | **22.3±0.2** | **61.4±0.8** | **16.6±0.2** | **4.98±0.48** | **13.53±0.84** | **15.03±0.81** | **35.54±2.10** |

Table 8: Mean and standard deviation on the Flickr30k Entities **test set**. All results are averaged **across five runs**.

| Method | Captioning Evaluation B@1 | B@4 | M | C | S | Grounding Evaluation $F1_{all}$ | $F1_{loc}$ | $F1_{all\_per\_sent}$ | $F1_{loc\_per\_sent}$ |
|---|---|---|---|---|---|---|---|---|---|
| Baseline | 23.2±0.5 | 2.22±0.2 | 10.8±0.3 | 45.9±1.5 | **15.1±0.2** | 3.75±0.16 | 12.00±0.76 | 9.41±0.26 | 31.68±0.93 |
| Cyclical | **23.7±0.13** | **2.45±0.1** | **11.1±0.1** | **46.4±0.6** | 14.8±0.2 | **4.71±0.41** | **15.84±1.56** | **11.73±0.22** | **41.56±0.75** |

Table 9: Mean and standard deviation on the ActivityNet-Entities **val set**. All results are averaged **across five runs**.

the cyclical training regimen is around 75 epochs. The total training time takes about 1 day.

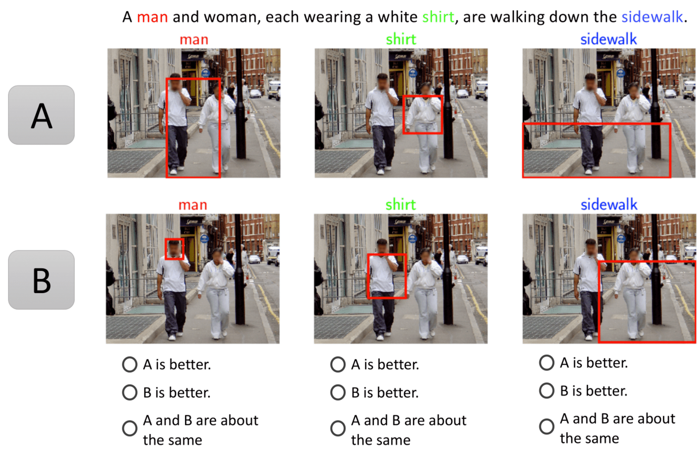


Fig. 2: Demonstration of our human evaluation study on grounding. Each human subject is required to rate which method (A or B) has a better grounding on each highlighted word.


Fig. 3: *A group of men in white uniforms are standing in a field with a crowd watching.* We can see that our proposed method attends to the sensible image regions for generating visually-groundable words, e.g., *man*, *uniforms*, *field*, and *crowd*. Interestingly, when generating *standing*, the model pays its attention on the image region with a foot on the ground.

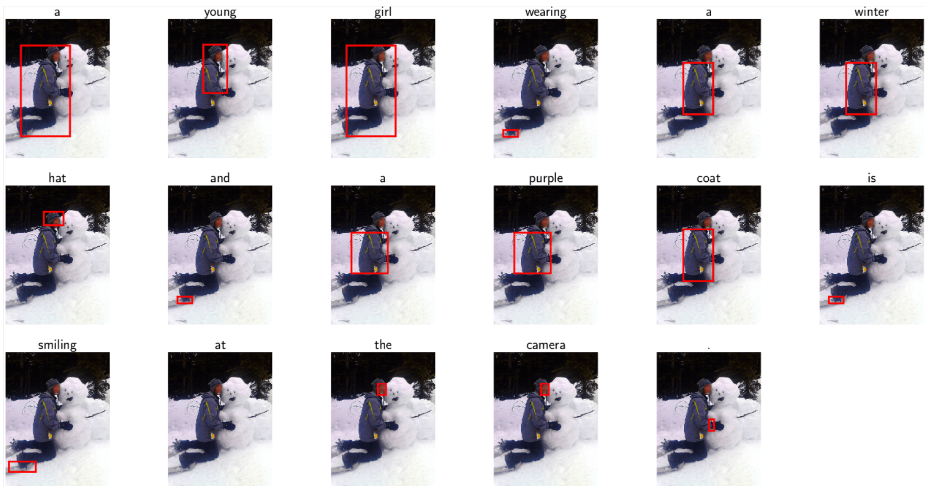


Fig. 4: *A young girl wearing a winter hat and a purple coat is smiling at the camera.* The proposed method is able to select the corresponding image regions to generate *girl*, *hat*, and *coat* correctly. We have also observed that the model tends to localize the person's face when generating *camera*.

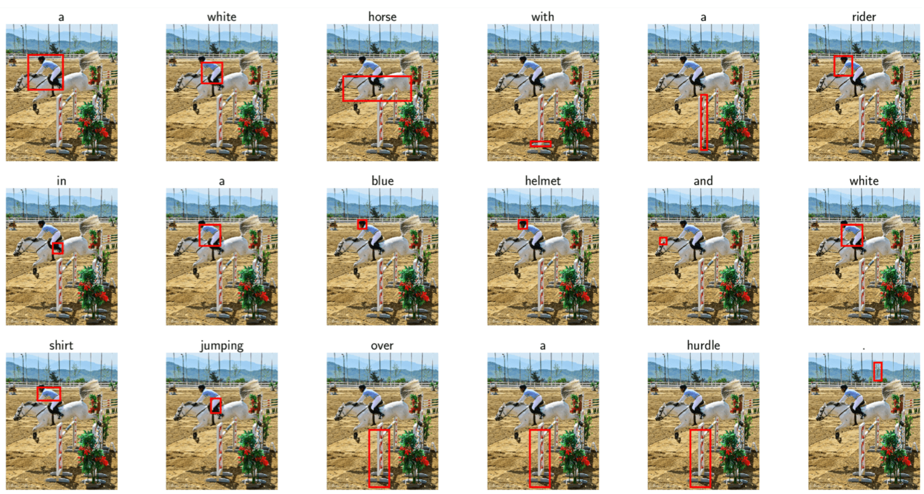


Fig. 5: *A white horse with a rider in a blue helmet and white shirt jumping over a hurdle.* While the model is able to correctly locate objects such as *horse*, *rider*, *helmet*, *shirt*, and *hurdle*, it mistakenly describes the rider as wearing a blue helmet, while it's actually black, and with white shirt while it's blue.

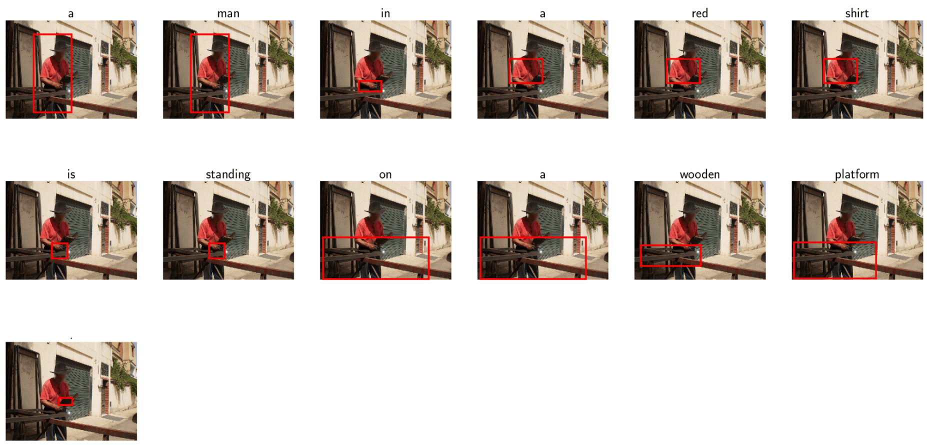


Fig. 6: *A man in a red shirt is standing on a wooden platform.* Our method correctly attends on the correct regions for generating *man*, *shirt*, and *platform*.


Fig. 7: *A man in a yellow jacket and blue helmet riding a bike.* The proposed method correctly generates a descriptive sentence while precisely attending to the image regions for each visually-groundable words: *man*, *jacket*, *helmet*, and *bike*.

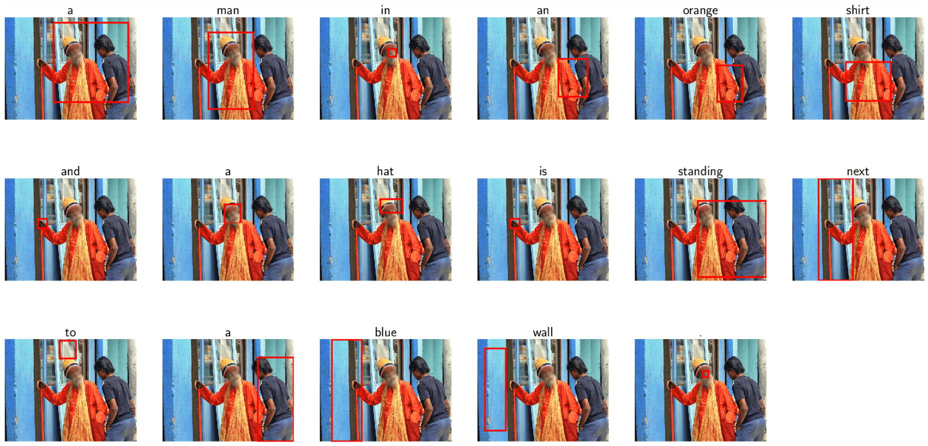


Fig. 8: *A man in an orange shirt and a hat is standing next to a blue wall.* While our method is able to ground the generated sentence on the objects like: *man*, *shirt*, *hat*, and *wall* , it completely ignores the person standing next to the man in the orange cloth.

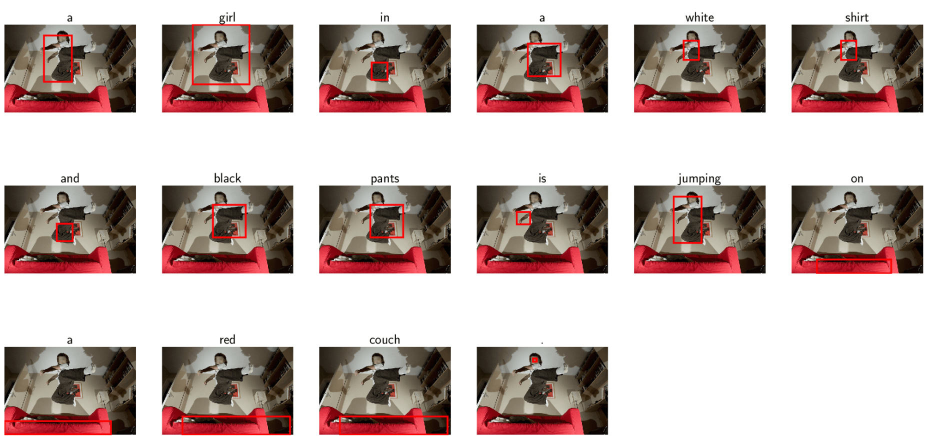


Fig. 9: *A girl in a white shirt and black pants is jumping on a red couch.* Our method is able to ground the generated descriptive sentence with the correct grounding on: *girl*, *shirt*, *pants*, and *couch*.


Fig. 10: *A man in a blue robe walks down a cobblestone street.* Our method grounds the visually-relevant words like: *man*, *robe*, and *street*. We can also see that it is able to locate the foot on ground for *walks*.

## References


1. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Proceedings of the International Conference on Learning Representations (ICLR) (2015) 5
2. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International Journal of Computer Vision **123**(1), 32–73 (2017) 4
3. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. In: Advances in Neural Information Processing Systems (NeurIPS). pp. 91–99 (2015) 4
4. Zhou, L., Kalantidis, Y., Chen, X., Corso, J.J., Rohrbach, M.: Grounded video description. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2019) 2, 4, 5
5. Zhou, L., Zhou, Y., Corso, J.J., Socher, R., Xiong, C.: End-to-end dense video captioning with masked transformer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 8739–8748 (2018) 5