Supplementary Material: Connecting What to Say With Where to Look by Modeling Human Attention Traces

1. Implementation Details of Downstream Task on COCO Captions

In the key boxes guided caption generation task defined by [1], at training time, we are provided a dense correspondence between caption and bounding boxes that associates every word in the caption with a bounding box detected by Faster-RCNN. If a word has no associated bounding box, [1] uses the average feature from all detected bounding boxes in this image as its corresponding visual representation. At inference time, the only input is the image with several ordered bounding boxes given by the user – no knowledge of which bounding box corresponds to which word(s) in the generated caption is assumed. [1] addressed this problem by proposing a specialized gate function to learn how to attend each word to the given boxes at inference time.

In contrast, our experiment proceeds under a slightly different setting: we are given the same information for both training and inference, i.e., only a sequence of bounding boxes for an image dense correspondence between boxes and words are not provided. We have the same inference setting as [1] but it is a more challenging training setting than [1], because we are provide sparser alignment of boxes and words during training.

Also note that this new task is different from the controlled caption generation task in our main paper. In controlled caption generation, the input is a meaningful smooth trace describing overall image content that often includes the relationship between objects and the background in the image, but our downstream task only utilizes several key box-object pairs.

To adapt such input into the same form as the setting where we deal with localized narratives dataset, we simply concatenate the given bounding boxes into a sequence, and pad $[0, 0, 1, 1, 1]$ if the length is less than what is needed (e.g., the length of sentence). In this way, the input becomes the same form (although the contained information is not exactly the same) that we used on the Localized Narratives experiments. Principally the specialized gate function proposed by [1] can also be added in our network, but we simply choose to use the same input format as our experiments on Localized Narratives because our goal here is to demonstrate the benefit of pre-training instead of trying to reproduce the setting in [1]. This results in a slightly different setting than that defined in [1]. As shown in the main paper, our pre-training on Localized Narratives brings clear gain under this new setting.

2. Layer Choice for Mirrored Transformer

We demonstrate the influence of the number of layers in our proposed mirrored transformer by varying the number of layers in the model trained with Task1 + Task2 + cycle$_b$ (cycle loss by permuting the trace within a mini batch). The results are shown in Table 1, which shows that two layers lead to better performance compared to one layer on both controlled trace generation (Task1) and controlled caption generation (Task2), while using three layers does not further improve results.

3. Influence of λ in Joint Training

We use the joint training of Task1 and Task2 as an example to show the influence of different λ values (defined in Eq. (3) in our main paper). The results are shown in Table 2. We can see that the performance of Task2 (controlled caption generation) remains relatively stable across different λ, while the performance of Task1 (controlled trace generation) improves when Task1 has a larger weight compared with Task2. In the experiments of the main paper, all values of λ are chosen from $\{1.0, 0.5, 0.3, 0.1, 0.0\}$ according to specific experiment settings and the performance on a subset of 5000 images from COCO2017 Training set (which we use to tune the hyperparameters).

4. More Qualitative Results

This section shows more qualitative results and analysis, for both success and failure cases of the model. In each subsection, the failure cases are ordered in descending order of subtlety (i.e., most obvious to most subtle).

#layers	BLEU-1	BLEU-4	METEOR	ROUGE$_L$	CIDEr	SPICE	LBM ($k=0$)	LBM ($k=1$)
1	0.598	0.286	0.258	0.479	1.407	0.313	0.166	0.155
2	**0.607**	**0.292**	**0.263**	**0.487**	**1.485**	**0.317**	**0.163**	**0.154**
3	0.604	0.289	0.261	0.485	1.444	0.314	0.197	0.191

Table 1. Influence of the number of layers. The model was trained on Task1 + Task2 + cycle$_b$, and evaluated on Task1 (LBM metric) and Task2 (other metrics in this table) respectively. Note: smaller values of LBM are better. The evaluation is performed on COCO2017 Validation set.

λ_2	BLEU-1	BLEU-4	METEOR	ROUGE$_L$	CIDEr	SPICE	LBM ($k=0$)	LBM ($k=1$)
1.0	0.589	0.272	0.254	0.472	1.346	0.306	0.187	0.178
0.5	**0.595**	0.278	**0.256**	**0.476**	1.368	**0.312**	0.179	0.169
0.3	0.586	0.272	0.252	0.470	1.329	0.307	0.169	0.157
0.1	**0.595**	**0.279**	**0.256**	0.474	**1.375**	0.310	**0.161**	**0.150**

Table 2. Results on different λ_2 when $\lambda_1 = 1$ for joint training of Task1 and Task2 (λ defined in eq. (3) in our main paper). The model was trained on Task1 + Task2, and evaluated on Task1 (LBM metric) and Task2 (other metrics in this table) respectively. Note: smaller values of LBM are better. The evaluation is done on COCO2017 Validation set.

4.1. Controlled Trace Generation

Below, we describe some successful instances and common failure cases of the model on the *controlled trace generation* task. Examples are all in Fig. 1, on the **left** column.

4.1.1 Successes

Correct object localization and spatial extent. The model successfully localizes the referred to objects and identifies their full spatial extents. For example, see row 1 (the animals and trees), row 2 (the woman, her hat, the stick), and row 3 (the tennis court, the woman, the racket, and the "adidas" text).

Recognition of directionality. The model attends to direction words in the input caption, such as "right" and "left." For example, in row 1, the caption specifies that there is a zebra and a giraffe in the "**right** side of the image" and the predicted trace correctly localizes these animals, even though there are other zebras and giraffes in the image that are described earlier in the caption. In row 2, where the caption begins "In the image in middle ...", the model quickly narrows in on the middle of the image (the red bounding boxes), rather than localizing the entire image. In row 3, the model correctly localizes the "**back**" of the image when the caption refers to this area.

Adapting to errors in the input caption. The model is able to adapt to errors in the input caption, such as spelling errors and incorrect object classifications. For example, in row 3, the model successfully localizes the "adidas" text when the caption reads "hoarding" (perhaps the annotator meant "heading"), and also localizes the tennis racket, even though it is referred to as a "bat."

4.1.2 Failure cases

False negatives in the trace. A ground-truth object or concept is not localized by the predicted trace. For example, see row 6. The predicted trace does not include the baseball bats and grass. This could be due to two reasons: (1) the model recognizes the bats and fence as relevant, but incorrectly localizes them, or (2) the model is focusing on larger and more visually dominant false positive objects in the image (such as the red columns in this image), and neglects the bats and fence from the trace.

Incorrect object spatial extent. A predicted region has the correct localization (e.g., the bounding box is positioned in the center of the referred-to object), but its spatial extent does not cover the full object. For example, see row 4: the man and the sheep are correctly localized, but their predicted spatial extents are too small. Another example is row 7: the predicted box for "sand" has good precision, in that it only localizes sand, but it does not cover the full spatial extent of the sand.

Poor region differentiation. Predicted regions referring to different objects/regions are correctly localized and have reasonable spatial extents, but these regions are not specific to each object. For example, see row 5. The duck and the water are correctly localized, but there is no way to differentiate these regions (since they all cover the same area around the duck). Ideally, the model would predict a tight box around the duck and a much larger box covering the entire water region.

4.2. Controlled Caption Generation

In this section, we describe some successful instances and common failure cases of the model on the *controlled caption generation* task. Examples are all in Fig. 1, on the **right** column.

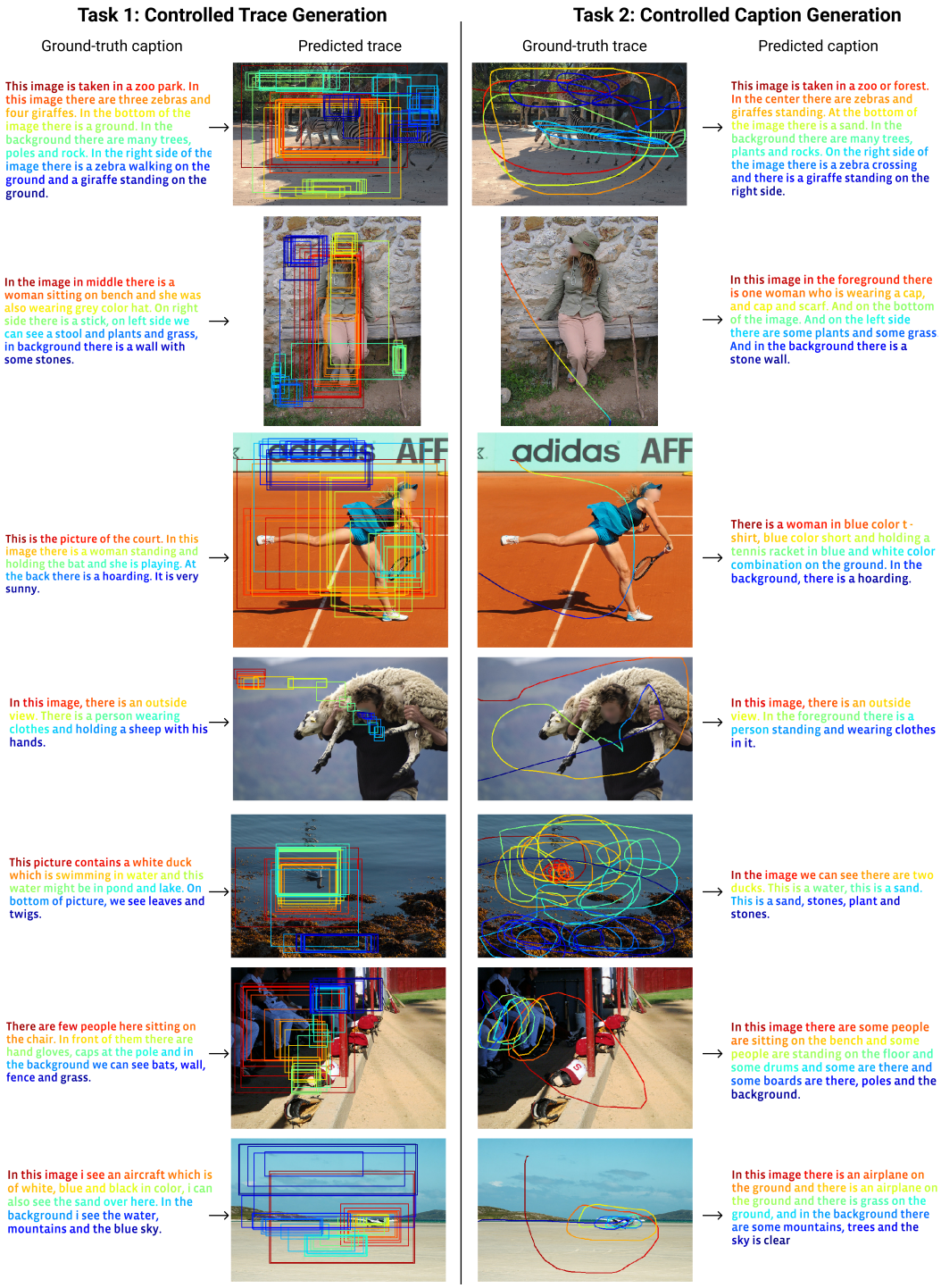

Figure 1. Qualitative results and selected failure cases on Tasks 1 and 2. The model was trained on Task1 + Task2 + cycle$_b$.

4.2.1 Successes

All major image components are described. The model successfully describes all the objects and "stuff" in the image, as compared to the ground-truth caption. See rows 1-3 for examples.

Caption is grammatical and tells a story. The predicted caption uses proper grammar, introduces the image (e.g., "In this image, there is ..."), and moves around the image describing different regions and objects. See rows 1-3 for examples.

Caption includes directionality. The model uses directions to refer to specific regions of the image it is describing. For example, see row 1 (the zebras and giraffes on the "**center** and "**right side**" of the image) and row 2 (the woman in the "**foreground**", the plants and grass on the "**left side**", and the stone wall in the "**background**".

Adapting to poor input traces. The model can output a rich caption, even given a poor input trace where the annotator drew a trace that is uncorrelated with the ground-truth caption. See rows 2 and 3 for examples.

4.2.2 Failure cases

False negative objects in the caption. Here, the caption omits mentioning a visually significant object or region that is specified by the human-provided trace. For example, see row 1 (the predicted caption neglects to mention the sheep) or row 7 (the model misses the sand).

False positive objects in the caption. Here, the caption hallucinates objects that are not present in the image. For example, see row 6: the model incorrectly describes the image as containing "drums."

Incorrect object counts. In this failure case, the model predicts the wrong number of objects that are present in the image. For example, see row 5: the model incorrectly predicts that there are "two ducks," rather than just one.

Object repetition. Here, the model correctly identifies an object in the image (that only has one instance), but mentions it multiple times. For example, see row 7: the model mentions the airplane twice ("there is an airplane on the ground and there is an airplane on the ground"), even though the image contains only instance of "airplane."

Grammar errors. This is a fairly common error, where the contents of the caption are correct, but the model uses incorrect grammar. For example, see row 5 ("this is a water, this is a sand"), and row 6 ("and some there are and some boards are there"). In many cases, the ground-truth captions have incorrect grammar (for example, see Fig. 1, left side, row 5), which could cause the model to learn and internalize these errors.

4.3. Joint Caption And Trace Generation

In this section, we describe some successful instances and common failure cases of the model on the *joint caption and trace generation* task. The model may also experience the issues described in Sections 4.1 and 4.2, but we focus on the errors specific to Task 3. Examples are all in Fig. 2.

4.3.1 Successes

Successful examples follow all the qualities of Task 1 and 2 (correct object localizations and spatial extents in the predicted trace, and precise, descriptive, and comprehensive predicted captions). They also have good alignment between the boxes in the predicted trace and the words in the caption. See rows 1 and 2 for examples.

4.3.2 Failure cases

Unaligned caption and trace. In this failure case, the model predicts a much longer trace than what is reasonable for the caption. For example, see row 3: the predicted trace has length 100, while the caption only has 18 words.

Caption cuts off due to maximum length constraint. In this case, the model is forced to stop predicting words because it hits the maximum caption length requirement (in this paper, this value is 100). This error usually happens in images that contain many distinct objects, because it is challenging for the model to group objects for conciseness. See row 4 for an example: the caption ends in the middle of a sentence ("On the ground there is a giraffe"). One solution would be to enforce the maximum caption length *and* require the caption to terminate with a full sentence (i.e., ending with a period), rather than allowing sentence fragments.

Figure 2. Qualitative results and selected failure cases on Task 3. The model was trained on Task3 + random mask.

References

[1] Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Show, control and tell: A framework for generating controllable and grounded captions. In *CVPR*, 2019. 1