6 Supplemental material

6.1 Synthetic Data Generation

For each novel object, we use the top similar candidate objects that were identified by FDM-net [8]. For e.g., images with 'cup', 'glass', or 'vase' are considered as candidates to generate synthetic data for novel object 'bottle'. We exclude four of these candidate object classes ('stove', 'bread', 'box', and 'racquet') as they do not have bounding box annotations in MS COCO dataset. We show our final list in Table 4.

6.2 Baseline Results

In order to get a good baseline model for Step I of our PS-NOC approach, we train the Up-Down model [5] using only the in-domain fully paired data and compare the validation split CIDEr against the baseline results of PS3 [4], which also uses the Up-Down model. With image features extracted using our ResNet-101 backbone based Faster R-CNN, our Up-Down model baseline under-performed compared to PS3 baseline. We were however able to achieve comparable baseline results by using ResNeXt-152 backbone based image features, as shown in Row 1 vs. Row 2 of Table 5. We also train another baseline model on the complete MS COCO dataset (i.e. 82,783 image-caption pairs), and compare our results against the corresponding baseline results in PS3; these are shown in Row 3 vs. Row 4 of Table 5.

Both these results show that our primary metrics of comparison, CIDEr and F1-score on out-of-domain test set, are either comparable or slightly lower than the baseline results in PS3, and that the ResNeXt-152 image features do not provide us an unfair advantage during comparison of PS-NOC with previous works or other state-of-the-art results.

6.3 Implementation Details

Model Architecture: We use the popular Bottom-Up Top-Down (Up-Down) captioning model [5] that is based on encoder-decoder neural architecture. The encoder consists of a Faster R-CNN [22] object detector that is trained on Visual Genome dataset [16]. For each image I, a set of feature vectors V associated with salient bounding boxes in the image are extracted using Faster R-CNN. At each timestep t (i.e., word in the caption), the decoder, which consists of a two-layer Long Short-Term Memory (LSTM) network, uses V and previous word y_{t-1} in the caption to compute the conditional probability of the next word $p_\theta(y_t \,|\, y_{1:t-1}, I)$ for all the words in the word vocabulary, where θ are the model parameters. We use Faster R-CNN with ResNext-152 as backbone. In the Supplemental material, we give details on the reason for choosing this backbone and why it does not provide us an unfair advantage during comparison with previous works. For language embeddings, we follow the same approach as PS3 [4], and add pre-trained word embeddings to the input and output layers of the decoder. However, we simply use GloVe embeddings [20] and freeze them throughout our model training.

Model Checkpoints: During Steps II and III of our training, we use the validation set scores to save the model checkpoint with the highest out-of-domain CIDEr, while also considering F1-score. Specifically, we allow a slight drop (1 point) in out-of-domain CIDEr if the corresponding F1-score is higher.

Table 4: The top similar candidate objects in fully paired data for each novel object class. This list is filtered down from the list in [8]

bottle	bus	couch	microwave	pizza	racket	suitcase	zebra
cup	truck	chair	refrigerator	sandwich	bat	handbag	giraffe
wine glass	car	bed	toaster	cake	frisbee	backpack	elephant
vase	train	bench					cow

Table 5: Comparison of our PS-NOC baseline Step I results against PS3 baseline results on validation split. Using ResNeXt-152 image features does not give us an unfair advantage while comparing our PS-NOC results with previous works or state-of-the-art results, since our out-of-domain CIDEr and F1 scores are comparable to PS3 baseline scores. We later fine-tune the model from Row 2 here using PS-NOC Steps II and III.

			Out-of-domain				In-domain		
Row	Approach	Training data	S	M	C	F1	S	M	C
1	PS3	In-domain (i.e., fully paired only)	14.4	22.1	69.5	0.0	19.9	26.5	108.6
2	Ours	In-domain (i.e., fully paired only)	19.4	27.9	66.7	0.0	19.5	27.8	107.8
3	PS3	In-domain + Out-of-domain	20.1	27	111.5	69.0	20.0	26.7	109.5
4	Ours	In-domain + Out-of-domain	20.6	29.1	107.1	60.0	20.7	29	109.0

6.4 Caption Post-processing and Sol-2

We notice that using SCST during training sometimes results in captions that end with the words 'in', 'a', 'with', etc. To improve caption quality, we perform post-processing during SCST-F1 training and inference, and remove such words at the end of the caption. We refer to our approach without such post-processing as Sol-1, and the one with the post-processing as Sol-2. The complete list of these words is as follows: 'with' ,'in' ,'on', 'of', 'a', 'at', 'to', 'for', 'an', 'this', 'his', 'her', 'that', 'the'. We provided the results for both these solutions in Table 1. We show a few qualitative examples of these results in Figure 5.

We used a slightly different training schedule for PS-NOC Step III in our Sol-2, compared to our Sol-1. In our Sol-1, we fine-tuned the model using SCST-F1 for $N = 4$ rounds with 6,000 iterations per round, initial LR of 0.002 for first round and scale it by 0.8 across rounds. In our Sol-2, we fine-tuned the model using SCST-F1 for $N = 4$ rounds with 8,000 iterations per round, initial LR of 0.003 for first round and scale it by 0.6 across rounds.

6.5 Ablation Studies

We perform ablation study to evaluate the effectiveness of our pseudo-labeling approach (see Section 3.2), wherein we generate two pseudo-label captions per partially paired image, one caption using CBS and the other without CBS. We also study the impact of the volume of synthetic data used during our PS-NOC training on the captioning results.

In Table 6, we provide the test split results from these additional ablation studies. We denote the three Steps in our PS-NOC approach as I, II and III respectively. II denotes cross-entropy loss training using our entire synthetic data (i.e., $K = 2400$ images per novel object), while II(33%) denotes similar training done using only 33% of this synthetic data (i.e., $K = 800$ images per novel object). III denotes cross-entropy loss training done using our pseudo-labeling approach, while III(CBS) denotes similar cross-entropy loss training

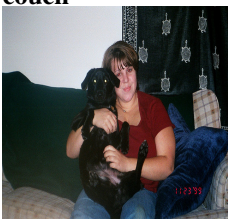
zebra

Baseline: A person sitting in a car with a large giraffe.
Sol-1: A **zebra** is standing in a car with a.
Sol-2: A **zebra** is standing in the seat of a car.

bus

Baseline: A police car parked on the side of the road.
Sol-1: A **bus** is parked on the side of a.
Sol-2: A **bus** is parked on a city street.

couch



Baseline: A woman is holding a black dog and a black dog.
Sol-1: A woman sitting on a **couch** with a black dog in.
Sol-2: A woman sitting on a **couch** holding a black dog.

microwave

Baseline: A cat is sitting on top of a toaster.
Sol-1: A cat laying on top of a **microwave** in a.
Sol-2: A cat laying on top of a **microwave**.

pizza

Baseline: A man sitting at a table with plates of food.
Sol-1: A man sitting at a table with a **pizza** and.
Sol-2: A man sitting at a table with a **pizza**.

racket

Baseline: A man is standing next to a tennis ball.
Sol-1: A man holding a tennis **racket** in front of a.
Sol-2: A man holding a tennis **racket** on a field.

suitcase

Baseline: A black and white cat is sitting in a car.
Sol-1: A black and white cat laying in a **suitcase** on.
Sol-2: A black and white cat laying in a **suitcase**.

bottle

Baseline: A brown dog laying next to a glass of wine.
Sol-1: A dog laying next to a **bottle** of wine on.
Sol-2: A dog laying next to a **bottle** of wine.

Figure 5: Examples of captions generated using our PS-NOC approach compared to the baseline model for the eight novel object classes. Our Sol-1 and Sol-2 both correctly include the novel object classes in the captions. Sol-2 includes caption post-processing during training and inference and has better caption quality than our Sol-1.

Table 6: Test split results of our ablation study showing the benefits of using i. Our pseudo-labeling approach, and ii. More synthetic data.

		Out-of-domain				In-domain		
Row	Training technique	S	M	C	F1	S	M	C
1	I + III(CBS)	19.9	28.1	92.7	60.8	19.6	27.8	105.9
2	I + III	19.4	28.0	86.9	60.9	19.1	27.6	105.5
3	I + II(33%) + III(CBS)	20.2	28.3	92.4	68.8	19.8	27.9	106.7
4	I + II(33%) + III	20.1	28.0	93.7	69.1	19.7	27.7	105.6
5	I + II + III(CBS)	20.0	28.3	94.4	74.1	19.6	27.6	105.0
6	I + II + III	20.1	28.3	96.5	74.0	19.7	27.8	105.0

but done using only the CBS generated pseudo-label caption. In these studies, we follow the same training schedule for Steps I and II as in Section 4.2. In Step III, we train the model for 15,000 iterations per round with an initial LR of 0.0025 for the first round and scale it by 0.5 across rounds. Table 6 demonstrates the following.

i. Our pseudo-labeling approach improves the out-of-domain caption quality, i.e., CIDEr scores, as seen in Row 3 vs. Row 4 and Row 5 vs. Row 6. This improvement is only seen if we use synthetic data (Rows 3-6) as proposed in our PS-NOC. (Rows 1 and 2 do not use synthetic data for training, as denoted by the missing II,)

ii. Using more synthetic data with our PS-NOC approach helps improve the out-of-domain scores as seen in Row 3 vs. Row 5 and Row 4 vs. Row 6.

Table 7: Test split results of our training Step III using different random seeds.

	Out-of-domain				In-domain		
Row	S	M	C	F1	S	M	C
1	20.8	28.0	103.8	85.9	20.5	27.7	110.9
2	20.9	27.7	103.0	85.9	20.5	27.6	110.5
3	20.7	27.7	103.2	86.8	20.4	27.3	110.1
4	20.8	27.9	103.0	84.5	20.3	27.5	109.7

Table 8: Using synthetic data in addition to complete MS COCO training data improves out-of-domain CIDEr and F1 scores on test split.

	Out-of-domain				In-domain		
Additional Synthetic Training data	S	M	C	F1	S	M	C
No	21.0	29.4	110.2	60.8	20.6	29.1	109.6
Yes	20.8	28.8	**111.3**	**65.6**	20.5	28.9	108.3

6.6 Ablation Studies on Reliability of Training Step III

We perform ablation study to evaluate the reliability of our training Step III by starting from the same Step II model checkpoint and training Step III using different random seeds. The test split results in Table 7 show that our training Step III results (using our Sol-2 approach) are consistent and reproducible using different random seeds. (Row 1 here is the result we show in Table 1, and Rows 2-4 are the other random seeds.)

6.7 Synthetic Data in Addition to Complete Data

We also check whether our synthetic data is useful to captioning models that have access to the complete training data, i.e., regular image captioning problem (not NOC).

We run a trial where we train the Up-Down model from scratch using both the complete MS COCO training data (i.e. 82,783 image-caption pairs) and our synthetic data (i.e. 18,974 image-caption pairs), and compare it against the model trained without using synthetic data. We train both these models for similar number of epochs, i.e., 49,000 and 40,000 iterations respectively to account for their dataset size difference, using a batch size of 100 and same LR schedule. The results in Table 8 on the test split show that using this additional synthetic data improves both out-of-domain CIDEr and F1-score by 1 and 5 points respectively. This encourages us to invest future research effort in generating such synthetic data for in-domain objects as well.

6.8 Limitations and Future Work

In this section, we discuss the limitations of our work and the potential future work to address them.

Object annotations for fully paired objects: During synthetic data generation, our approach requires that fully paired dataset includes bounding box annotations for objects, which is true for both the existing novel captioning datasets [1, 13] based on MS COCO. However, if such annotations are not available, this requirement can be relaxed by predicting

A blue plate holding a frosted cake and knife.
+ Novel Object: pizza
= A plate holding a pizza and knife.

A birthday cake has a fraction of itself cut and eaten.
+ Novel Object: pizza
= A pizza has a fraction of itself cut and eaten.

Figure 6: Examples of synthetic data generated using our approach. The first caption is accurate while the second has incorrect context ('pizza' is not cut or eaten).

these object bounding boxes using object detection models, and then using them to generate synthetic images.

Noise in synthetic data: There could be cases where ground truth object bounding box annotations are incorrect or partial, which would result in noisy synthetic images. There could also be cases where our simple caption processing heuristics fail, or the context of synthetic data may not be completely accurate, resulting in noisy synthetic captions. For e.g., 'a fraction of itself cut and eaten' is not accurate description of the synthetic 'pizza' in Figure 6. However, our empirical results (Row 1 vs. 3 in Table 3) show that synthetic data generated using such simple heuristics is also able to considerably improve the NOC model results. More sophisticated techniques to remove such noise could improve the captioning results further.

Other datasets and models: We focused our extensive evaluation only on Up-Down captioning model and *held-out* MS COCO dataset. Evaluation of PS-NOC on the newer Transformer based captioning models [14] and *no-caps* dataset [1] will be pursued as future work.