

# 1 Appendix

## 2 0.1 Data augmentation

3 Fig. 1 shows some examples of augmented MSCOCO images and captions. We perform image-  
 4 level augmentation to the input images of a Faster R-CNN (pretrained on the Visual Genome<sup>1</sup>) and  
 5 apply RoI-level augmentation to the bounding boxes/ROIs detected by Faster R-CNN. For sentence  
 6 augmentation, we use the transformer-based neural machine translation models [1] pretrained on  
 7 WMT'19<sup>2</sup> to perform back-translation. For ground-truth dependency trees, we parse each sentence  
 8 with the dependency parser provided by Stanza<sup>3</sup>.

Image (Original)	Sentences (Original)	Sentences (Back-Translation)	Image (HFlip)		
	A dog standing in the grass near a flying frisbee.	En-De-EN <i>A dog stands in the grass next to a flying Frisbee.</i> En-Ru-EN <i>A dog standing in the grass next to a flying Frisbee.</i>			
	A cute little dog running through a yard towards a frisbee.	En-De-EN <i>A cute little dog running through a yard towards a Frisbee.</i> En-Ru-EN <i>Cute little dog runs through yard to Frisbee.</i>			
	A dog looks at a Frisbee as it flies toward it.	En-De-EN <i>A dog looks at a Frisbee as it flies towards it.</i> En-Ru-EN <i>A dog looks at Frisbee as he flies towards her.</i>			
	A dog in a yard catching a Frisbee.	En-De-EN <i>A dog on a farm catches a Frisbee.</i> En-Ru-EN <i>Dog catches Frisbee in the yard.</i>			
	A dog chasing after a purple frisbee on top of a green lawn.	En-De-EN <i>A dog chases a purple frisbee on a green lawn.</i> En-Ru-EN <i>A dog chases a purple Frisbee over a green lily.</i>			
RoI Horizontal Flip	RoI Jitter [0.9, 1.1]	RoI Jitter [0.8, 1.2]	RoI Jitter [0.8, 1.2]	RoI Jitter [0.9, 1.1]	RoI Horizontal Flip
					
RoI Rotate 90	RoI Rotate 180	RoI Rotate 270	RoI Rotate 270	RoI Rotate 180	RoI Rotate 90
					

Image (Original)	Sentences (Original)	Sentences (Back-Translation)	Image (HFlip)		
	A man and two women walking their dogs and hiking in the woods.	En-De-EN <i>A man and two women walk their dogs through the woods.</i> En-Ru-EN <i>A dog standing in the grass next to a flying Frisbee.</i>			
	Three dogs are huddled together with their owners.	En-De-EN <i>Three dogs are crammed together with their owners.</i> En-Ru-EN <i>Three dogs walk with their owners.</i>			
	A group of people in the woods with dogs.	En-De-EN <i>A group of people in the forest with dogs.</i> En-Ru-EN <i>A group of masked men with dogs.</i>			
	Three hikers walking in the wilderness with three dogs.	En-De-EN <i>Three hikers take three dogs for a walk in the wilderness.</i> En-Ru-EN <i>Three hikers walking in the wild with three dogs.</i>			
	Three people standing in a wooded area walking three dogs.	En-De-EN <i>Three people stand in a wooded area and run three dogs.</i> En-Ru-EN <i>Three people stand in the forest, walking three dogs</i>			
RoI Horizontal Flip	RoI Jitter [0.9, 1.1]	RoI Jitter [0.8, 1.2]	RoI Jitter [0.8, 1.2]	RoI Jitter [0.9, 1.1]	RoI Horizontal Flip
					
RoI Rotate 90	RoI Rotate 180	RoI Rotate 270	RoI Rotate 270	RoI Rotate 180	RoI Rotate 90
					

Figure 1: Examples of augmented MSCOCO images and captions. For each augmented image, we show the object labels at the centers of respective bounding box for a better visualization. We apply per-category non-maximum suppression to the raw bounding boxes detected by Faster R-CNN.

<sup>1</sup><https://github.com/airsplay/py-bottom-up-attention>

<sup>2</sup><https://github.com/pytorch/fairseq>

<sup>3</sup><https://github.com/stanfordnlp/stanza>

## 9 0.2 Implementation details

### 10 0.2.1 Details of pretraining

11 Fig. 2 illustrates the visual-textual alignment mechanisms of the three variants of our proposed SSRP.  
 12 For  $\text{SSRP}_{\text{Cross}}$ , we take the final hidden state of [CLS] to predict whether the sentence matches  
 13 with the image semantically. For  $\text{SSRP}_{\text{Share}}$  and  $\text{SSRP}_{\text{Visual}}$ , since they do not have the bidirectional  
 14 cross-attention as in  $\text{SSRP}_{\text{Cross}}$ , we take  $\sum_i v_i/N_v$  as the additional input and concatenate it with  
 15  $w_{\text{CLS}}$  to generate the visual-textual alignment prediction using  $g_{\text{align}}(\cdot)$ .

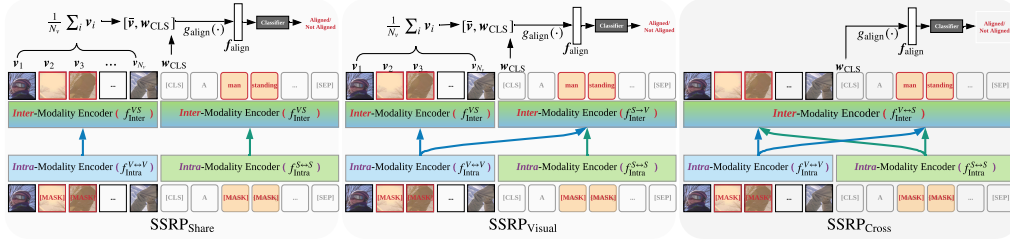


Figure 2: An illustration of the mechanism used for obtaining the visual-textual alignment representation  $f_{\text{align}}$  for each of the three SSRP variants. These three SSRP variants can be used to facilitate the fine-tuning for different downstream tasks. Note that,  $\text{SSRP}_{\text{Cross}}$  can only support visual-textual multi-modal downstream tasks such as VQA, while  $\text{SSRP}_{\text{Share}}$  and  $\text{SSRP}_{\text{Visual}}$  can support not only multi-modal downstream tasks but also single-modal visual tasks such as image captioning.

### 16 0.2.2 Details of NLVR2 fine-tuning

17 NLVR2 is a challenging visual reasoning task. It requires the model to determine whether the natural  
 18 language statement  $S$  is true about an image pair  $\langle I_i, I_j \rangle$ . During both fine-tuning and testing, we  
 19 feed alignment representations of the two images and the probed relationships to a binary classifier.  
 20 The predicted probability is computed as:

$$p(I_i, I_j, S) = \sigma(f_{\text{FC}}(f_{\text{p}}([q_i; q_j]))) \quad (\text{A.1})$$

$$q_k = f_{\text{q}}([f_{\text{align}}^k; f_{\text{vw}}([R_k^v; R_k^w])]), \quad k \in [i, j] \quad (\text{A.2})$$

21 where  $f_{\text{align}}^k$ ,  $R_k^v$ , and  $R_k^w$  are the outputs of  $\text{SSRP}(I_k, S)$ , and  $\sigma$  denotes the sigmoid activation  
 22 function. The nonlinear transformation functions  $f_{\text{vw}}$ ,  $f_{\text{q}}$ ,  $f_{\text{p}}$  and linear FC layer  $f_{\text{FC}}$  have learnable  
 23 weights.

24 For baseline models that do not consider relationships, the predicted probability is computed as:

$$p(I_i, I_j, S) = \sigma(f_{\text{FC}}(f_{\text{p}}([f_{\text{align}}^i; f_{\text{align}}^j]))) \quad (\text{A.3})$$

25 We fine-tune all models (including SSRP) with sigmoid binary cross-entropy loss.

### 26 0.2.3 Details of VQA/GQA fine-tuning

27 VQA requires the model to answer a natural language question  $Q$  related to an image  $I$ . We conduct  
 28 experiments on the VQA v2.0 dataset. We fine-tune our model on the train split using sigmoid  
 29 binary cross-entropy loss and evaluate it on the test-standard split. Note that VQA is based on  
 30 the MSCOCO image corpus, but the questions have never been seen by the model during training.  
 31 During fine-tuning, we feed the region features and given question into  $\text{SSRP}_{\text{Cross}}$ , and then output the  
 32 alignment representation and the probed relationships that are fed to a classifier for answer prediction:

$$p(I, Q) = \sigma(f_{\text{FC}}(f_{\text{p}}(\mathbf{q}))) \quad (\text{A.4})$$

$$\mathbf{q} = f_{\text{q}}([f_{\text{align}}; f_{\text{vw}}([R^v; R^w])]) \quad (\text{A.5})$$

33 where  $f_{\text{align}}$ ,  $R^v$ , and  $R^w$  are the outputs of  $\text{SSRP}_{\text{Cross}}$ , and  $\sigma$  denotes the sigmoid activation function.  
 34 The nonlinear transformation functions  $f_{\text{vw}}$ ,  $f_{\text{q}}$ ,  $f_{\text{p}}$  and linear FC layer  $f_{\text{FC}}$  have learnable weights.

### 35 0.2.4 Details of image captioning

36 For image captioning, we use only the image branch of  $\text{SSRP}_{\text{Visual}}$ , and feed the unmasked image  
 37 features into  $\text{SSRP}_{\text{Visual}}$ . For each input image, we first extract the contextualized visual representation  
 38  $\mathbf{v}_{1:N_v}$  and the implicit visual relationships  $\mathbf{R}^v$  from the pretrained  $\text{SSRP}_{\text{Visual}}$ . The inputs to the image  
 39 captioning model are the refined object features  $\mathbf{v}_{1:N_v}$  and probed relationships  $\mathbf{R}^v$ . We treat  $\mathbf{R}^v$  as  
 40 a global representation for the image. We set the number of hidden units of each LSTM to 1000, the  
 41 number of hidden units in the attention layer to 512. We first optimize the model on one Tesla V100  
 42 GPU using cross-entropy loss, with an initial learning rate of  $5e-4$ , a momentum parameter of 0.9,  
 43 and a batch size of 100 for 40 epochs. After that, we further train the model to optimize it directly for  
 44 CIDEr score [2] for another 100 epochs. During testing, we adopt beam search with a beam size of 5.  
 45 We apply the same training and testing settings for Up-Down (Our Impl.) and  $\text{SSRP}_{\text{Visual}}$ .

### 46 0.2.5 Details of image retrieval

47 For image retrieval, we also feed the unmasked image features into  $\text{SSRP}_{\text{Visual}}$  and obtain the refined  
 48 contextualized visual representations along with the implicit visual relationships. We conduct the  
 49 retrieval experiment on MSCOCO validation set. We randomly sample the query images and retrieve  
 50 the top images according to their cosine similarities against the queries.

51 We compare two kinds of methods: one that uses contextualized visual representations  $\mathbf{v}_{1:N_v}$ , and  
 52 another one that uses both contextualized visual representations  $\mathbf{v}_{1:N_v}$  and implicit visual relationships  
 53  $\mathbf{R}^v$ . For ‘Obj. + Rel.’ approach, we use the relationship-enhanced visual features obtained with  
 54  $\frac{1}{N_v} \sum_i \frac{1}{N_v} \sum_k \mathbf{v}_i d_{B_v}(\mathbf{v}_i, \mathbf{v}_k)^2$ . For ‘Obj.’ approach, we simply average the contextualized object  
 55 features with  $\frac{1}{N_v} \sum_i \mathbf{v}_i$ . Fig. 3 shows the pipeline for the image retrieval task.

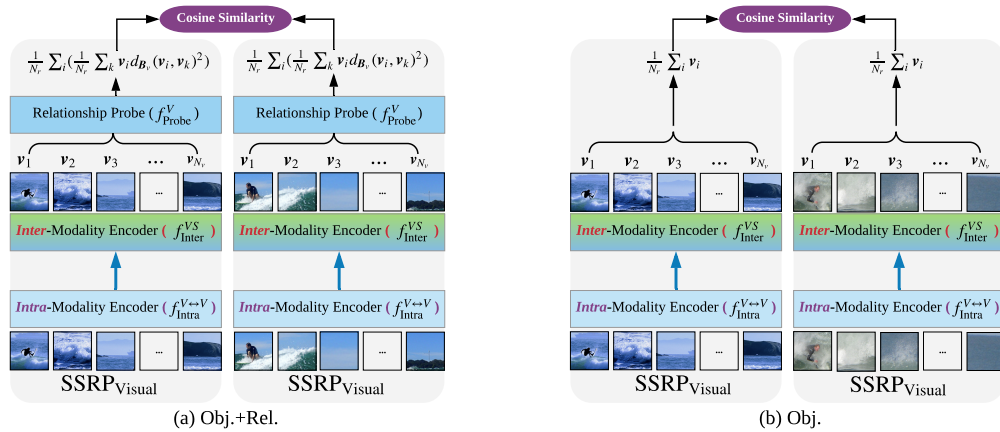


Figure 3: Illustrations of the two image retrieval methods mentioned in our paper.

56 **0.3 Extra examples**

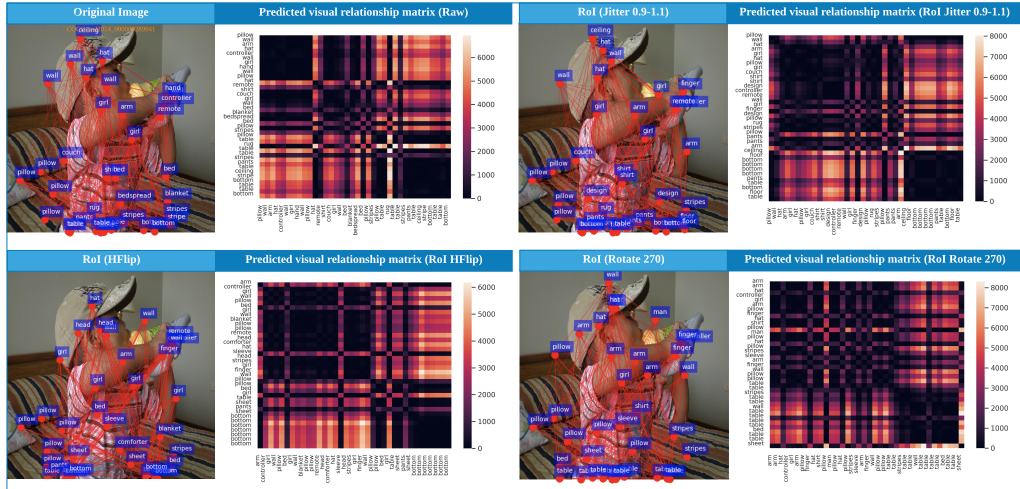


Figure 4: Examples of generated relationships for different augmented images. Darker colors indicate closer visual relationships, while lighter colors indicate farther visual relationships.

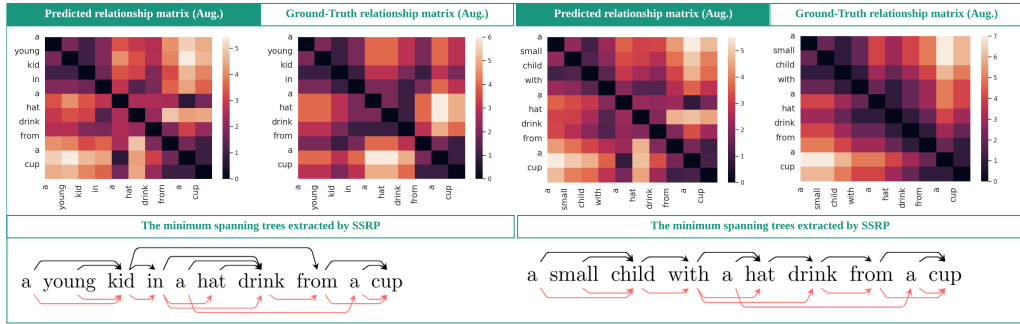


Figure 5: Example of generated relationships for different augmented sentences. Bottom row shows the minimum spanning trees. Black edges are the ground-truth parse; red are predicted by SSRP<sub>Cross</sub>.

57 **References**

58 [1] Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. Facebook  
 59 fair’s wmt19 news translation task submission. *arXiv preprint arXiv:1907.06616*, 2019.

60 [2] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-  
 61 critical sequence training for image captioning. In *CVPR*, 2017.