Show, Control and Tell: A Framework for Generating Controllable and Grounded Captions

Supplementary Material

Marcella Cornia Lorenzo Baraldi Rita Cucchiara
University of Modena and Reggio Emilia
{name.surname}@unimore.it

1. Sorting network

We provide additional details on the architecture and training strategy of the sorting network. For the ease of the reader, a schema is reported in Fig. 1. Given a scrambled sequence of N region sets, each region is encoded through a fully connected network which returns a N-dimensional descriptor. The fully connected network employs visual, textual and geometric features: the Faster R-CNN vector of the detection (2048-d), the GloVe embedding of the region class (300-d) and the normalized position and size of the bounding-box (4-d). The visual vector is processed by two layers (512-d, 128-d), while the textual feature is processed by a single layer (128-d). The outputs of the visual and textual branches are then concatenated with the geometric features and fed through another fully connected layer (256-d). A final layer produces the resulting N-dimensional descriptors. All layers have ReLU activations, except for the last fully-connected which has a $\tanh$ activation. In case the region set contains more than one detection, we average-pool the resulting N-dimensional descriptors to obtain a single feature vector for a region set.

Once the feature vectors of the scrambled sequence are concatenated, we get a $N \times N$ matrix, which is then converted into a "soft" permutation matrix $\boldsymbol{P}$ through the Sinkhorn operator. The operator processes a N-dimensional square matrix $\boldsymbol{X}$ by applying L consecutive row-wise and column-wise normalization, as follows:

$$S^0(\boldsymbol{X}) = \exp(\boldsymbol{X}) \tag{1}$$

$$S^l(\boldsymbol{X}) = \mathcal{T}c(\mathcal{T}_r(S^{l-1}(\boldsymbol{X}))) \tag{2}$$

$$\boldsymbol{P} := S^L(\boldsymbol{X}) \tag{3}$$

where $\mathcal{T}_r(\boldsymbol{X}) = \boldsymbol{X} \oslash (\boldsymbol{X}\mathbf{1}_N\mathbf{1}_N^T)$, and $\mathcal{T}_c(\boldsymbol{X}) = \boldsymbol{X} \oslash (\mathbf{1}_N\mathbf{1}_N^T\boldsymbol{X})$ are the row-wise and column-wise normalization operators, with $\oslash$ denoting element-wise division, $\mathbf{1}_N$ a column vector of N ones. At test time, once L normalizations ($L = 20$ in our experiments) have been performed, the resulting "soft" permutation matrix can be converted into a permutation matrix via the Hungarian algorithm [1].

At training time, instead, we measure the mean square error between the scrambled sequence and its reconstructed version obtained by applying the soft permutation matrix to the sorted ground-truth sequence $\boldsymbol{R}^*$, *i.e.* $\boldsymbol{P}^T\boldsymbol{R}^*$. On the implementation side, all tensors are appropriately masked to deal with variable-length sequences and sets. We set the maximum length of input scrambled sequences to 10.

In Table 1 we evaluate the quality of the rankings in

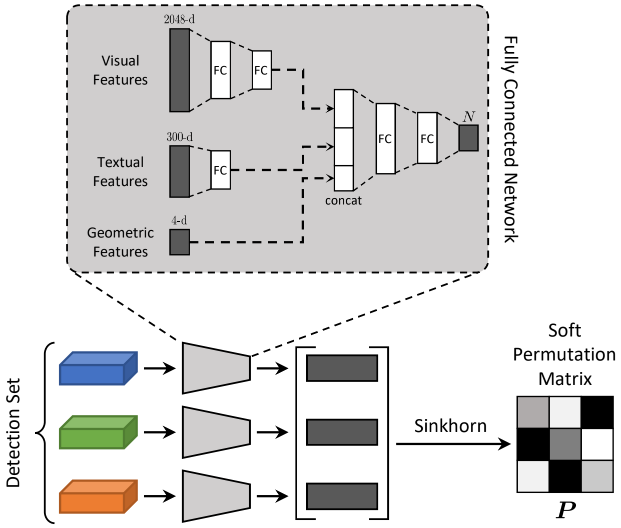
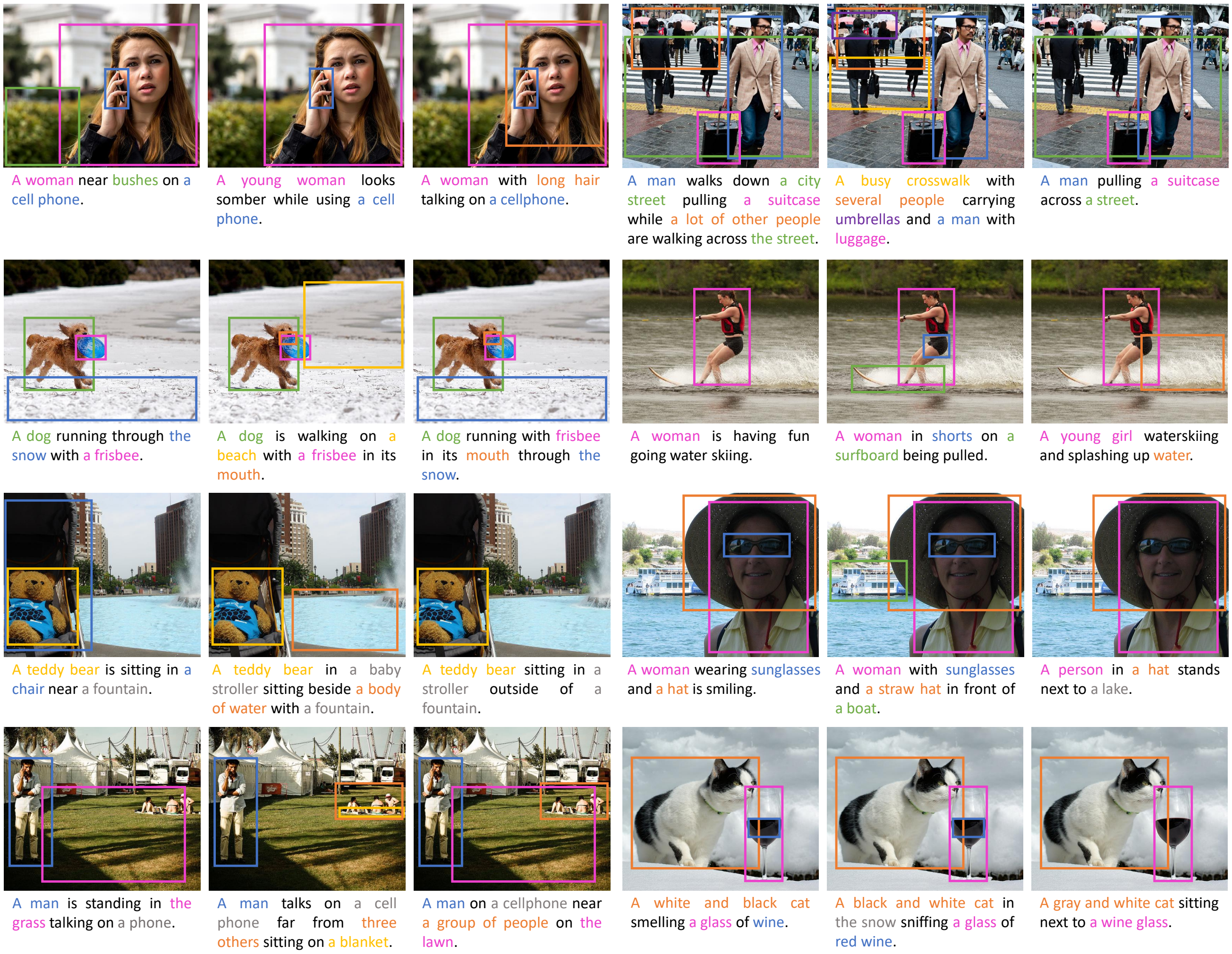


Figure 1: Schema of the sorting network.

	COCO Entities		Flickr Entities	
	Accuracy	Kendall's Tau	Accuracy	Kendall's Tau
Predefined local (det. prob.)	36.2%	0.145	40.0%	0.249
Predefined global (det. class)	59.1%	0.525	58.4%	0.565
SVM Rank	54.6%	0.448	49.5%	0.418
Sinkhorn Network	**67.1%**	**0.613**	**65.2%**	**0.633**

Table 1: Sorting network: experimental evaluation.



A woman near bushes on a cell phone.

A young woman looks somber while using a cell phone.

A woman with long hair talking on a cellphone.

A man walks down a city street pulling a suitcase while a lot of other people are walking across the street.

A busy crosswalk with several people carrying umbrellas and a man with luggage.

A man pulling a suitcase across a street.

A dog running through the snow with a frisbee.

A dog is walking on a beach with a frisbee in its mouth.

A dog running with frisbee in its mouth through the snow.

A woman is having fun going water skiing.

A woman in shorts on a surfboard being pulled.

A young girl waterskiing and splashing up water.

A teddy bear is sitting in a chair near a fountain.

A teddy bear in a baby stroller sitting beside a body of water with a fountain.

A teddy bear sitting in a stroller outside of a fountain.

A woman wearing sunglasses and a hat is smiling.

A woman with sunglasses and a straw hat in front of a boat.

A person in a hat stands next to a lake.

A man is standing in the grass talking on a phone.

A man talks on a cell phone far from three others sitting on a blanket.

A man on a cellphone near a group of people on the lawn.

A white and black cat smelling a glass of wine.

A black and white cat in the snow sniffing a glass of red wine.

A gray and white cat sitting next to a wine glass.

Figure 2: Additional sample captions and corresponding visual groundings from the COCO Entities dataset. Different colors show a correspondence between textual chunks and image regions. Gray color indicates noun chunks for which a visual grounding could not be found, either for missing detections or for errors in the noun-class association.

terms of accuracy (proportion of completely correct rankings) and Kendall's Tau (correlation between GT and predicted ranking, between -1 and 1). We compare with a predefined local ranking (sorting detections with their probability), a predefined global ranking based on detection classes, and compare the Sinkorn network with a SVM Rank model trained on the same features. As it can be seen, the Sinkhorn network performs better than other baselines and can generate accurate rankings.

2. Training details

We used a weight of 0.2 for the word loss and 0.8 for the two chunk-level terms in Eq. 11. To train both the captioning model and the sorting network, we use the Adam optimizer with an initial learning rate of 5×10^{-4} decreased by a factor of 0.8 every epoch. For the captioning model, we run the reinforcement learning training with a fixed learning rate of 5×10^{-5}. We use a batch size of 100 for all our experiments. During caption decoding, we employ for all experiments the beam search strategy with a beam size of 5: similarly to what has been done when training with Reinforcement Learning, we sample from both output distribution to select the most probable sequence of actions. We use early stopping on validation CIDEr for the captioning network, and validation accuracy of the predicted permutations for the sorting network.

3. The COCO Entities dataset

In Fig. 2, we report additional examples of the semi-automatic annotation procedure used to collect COCO Entities. As in the main paper, we use different colors to visualize the correspondences between noun chunks and image regions. For the ease of visualization, we display a single region for chunk, even though multiple associations are possible. In this case, the region set would contain more than one element.

In the last two rows, we also report samples in which at least one noun chunk could not be assigned to any detection. Recall that in this case, at training time, we use the most probable detections of the image and let the adaptive attention mechanism learn the corresponding association: we found that this procedure, overall, increases the final accuracy of the network rather than feeding empty region sets. Captions with missing associations are dropped in validation and testing.

4. Additional experimental results

Tables 2, 3 and 4 report additional experimental results which have not been reported in the main paper for space constraints. In particular, Table 2 integrates Table 3 of the main paper by evaluating the controllability via a sequence of region sets on Flickr30K, when training with cross-entropy only, and when optimizing with CIDEr and CIDEr+NW. Analogously, Tables 3 and 4 analyze the controllabilty via a set of regions, on both Flickr30K and COCO Entities and with all training strategies.

We observe that the CIDEr+NW fine-tuning approach is effective on all settings, and that our model outperforms by a clear margin the baselines both when controlled via a sequence and when controlled by a scrambled set of regions, regardless of the careful choice of the baselines. The performance of the Controllable LSTM baseline is constantly significantly lower than that of the Controllable Up-Down, thus indicating both the importance of an attention mechanism and that of having a good representation of the control signal. The Controllable Up-Down baseline, however, shows lower performance when compared to our approach, in both sequence- and set-controlled scenarios.

5. Additional qualitative results

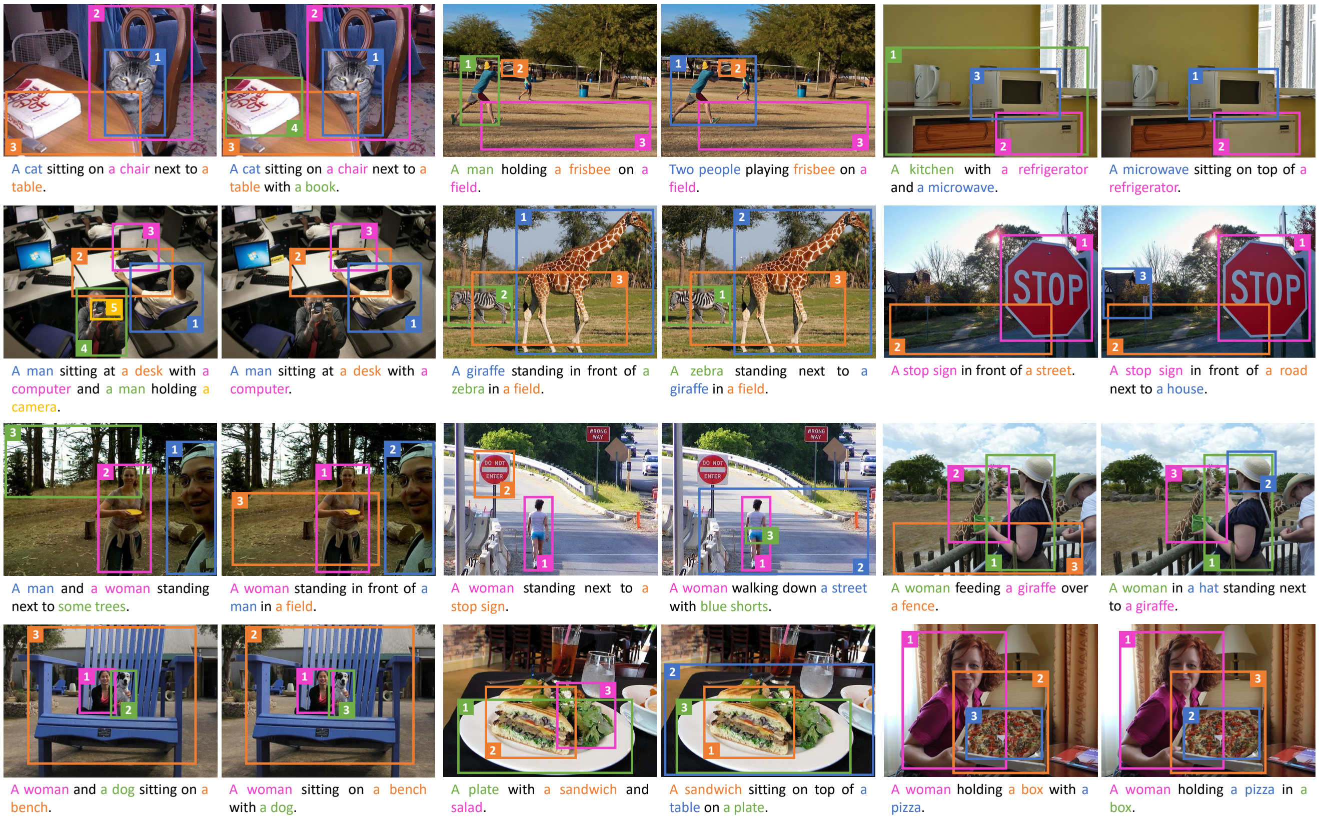
Finally, Fig. 3 and 4 report other qualitative results on COCO Entities. As in the main paper, the same image is reported multiple times with different control inputs: our method generates multiple captions for the same image, and can accurately follow the control input.

References

[1] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

	Cross-Entropy Loss						CIDEr Optimization						CIDEr + NW Optimization					
Method	B-4	M	R	C	S	NW	B-4	M	R	C	S	NW	B-4	M	R	C	S	NW
Controllable LSTM	6.5	12.0	29.6	40.4	15.7	0.078	6.7	12.1	30.0	45.5	15.8	0.079	6.5	12.6	30.2	43.5	15.8	0.124
Controllable Up-Down	10.1	15.2	34.9	69.2	21.6	**0.158**	10.1	14.8	35.0	69.3	21.2	0.148	10.4	15.2	35.2	69.5	21.7	0.190
Ours *w/* single sentinel	11.0	**15.5**	36.3	71.7	22.6	0.134	11.2	15.8	37.9	77.9	22.9	0.199	10.7	16.1	38.1	76.5	22.8	0.260
Ours *w/o* visual sentinel	10.8	14.9	35.4	69.3	22.2	0.142	11.1	15.5	36.8	75.0	22.2	0.197	11.1	15.5	37.2	74.7	22.4	0.244
Ours	**11.3**	15.4	**36.9**	**74.5**	**23.4**	0.152	**12.4**	**16.6**	**38.8**	**83.7**	**23.5**	**0.221**	**12.5**	**16.8**	**38.9**	**84.0**	**23.5**	**0.263**

Table 2: Controllability via a sequence of regions, on the test portion of Flickr30K Entities.


A cat sitting on a chair next to a table.

A cat sitting on a chair next to a table with a book.

A man holding a frisbee on a field.

Two people playing frisbee on a field.

A kitchen with a refrigerator and a microwave.

A microwave sitting on top of a refrigerator.

A man sitting at a desk with a computer and a man holding a camera.

A man sitting at a desk with a computer.

A giraffe standing in front of a zebra in a field.

A zebra standing next to a giraffe in a field.

A stop sign in front of a street.

A stop sign in front of a road next to a house.

A man and a woman standing next to some trees.

A woman standing in front of a man in a field.

A woman standing next to a stop sign.

A woman walking down a street with blue shorts.

A woman feeding a giraffe over a fence.

A woman in a hat standing next to a giraffe.

A woman and a dog sitting on a bench.

A woman sitting on a bench with a dog.

A plate with a sandwich and salad.

A sandwich sitting on top of a table on a plate.

A woman holding a box with a pizza.

A woman holding a pizza in a box.

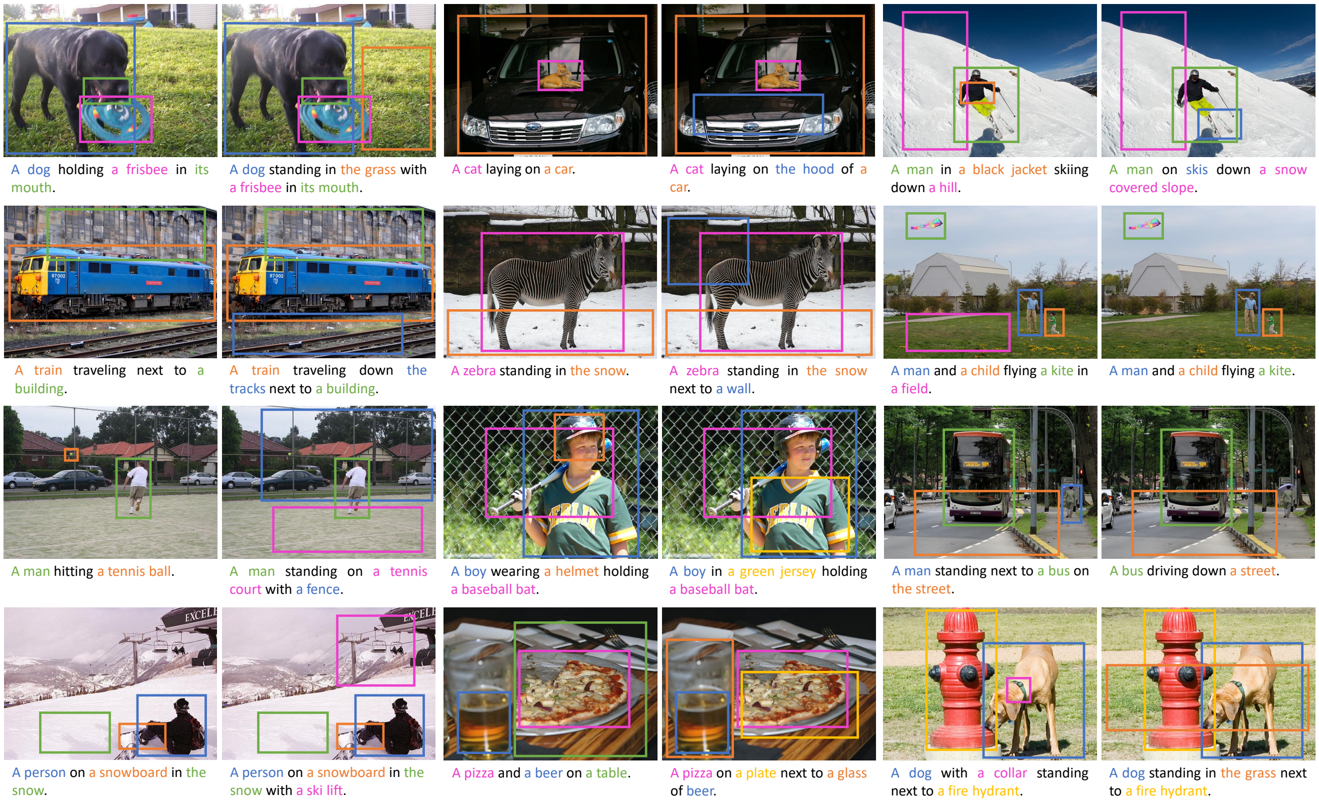
Figure 3: Additional sample results of controllability via a sequence of regions. Different colors and numbers show the control sequence and the associations between chunks and regions.

	Cross-Entropy Loss						CIDEr Optimization						CIDEr + NW Optimization					
Method	B-4	M	R	C	S	IoU	B-4	M	R	C	S	IoU	B-4	M	R	C	S	IoU
Controllable LSTM	11.5	18.1	38.5	105.8	27.1	60.7	12.9	18.9	40.9	122.0	28.2	62.0	12.9	19.3	41.3	123.4	28.7	0.642
Controllable Up-Down	17.5	23.0	46.9	160.6	38.8	69.2	17.7	22.9	47.3	167.6	38.7	69.4	**18.1**	23.6	48.4	170.5	40.4	71.6
Ours *w/* single sentinel	16.9	22.6	46.9	159.6	40.9	70.2	17.9	23.7	48.7	171.1	43.5	74.4	17.4	23.6	48.4	168.4	43.7	75.4
Ours *w/o* visual sentinel	**17.7**	23.1	47.9	166.6	**42.1**	71.3	18.1	23.7	48.9	172.5	43.3	74.2	17.6	23.4	48.5	168.9	43.6	75.3
Ours	**17.7**	**23.2**	**48.0**	**168.3**	**42.1**	**71.4**	**18.5**	**23.9**	**49.0**	**176.7**	**43.8**	**74.5**	18.0	**23.8**	**48.9**	**173.3**	**44.1**	**75.5**

Table 3: Controllability via a set of regions, on the test portion of COCO Entities.

	Cross-Entropy Loss						CIDEr Optimization						CIDEr + NW Optimization					
Method	B-4	M	R	C	S	IoU	B-4	M	R	C	S	IoU	B-4	M	R	C	S	IoU
Controllable LSTM	6.7	12.0	29.8	41.0	15.6	48.8	6.8	12.1	30.2	45.4	15.6	49.0	6.4	12.5	30.2	42.9	15.6	50.8
Controllable Up-Down	**10.1**	**15.2**	35.1	**68.8**	21.5	**53.6**	10.2	14.8	35.3	69.1	21.1	52.9	10.5	15.2	35.5	69.5	21.6	54.8
Ours *w/* single sentinel	**10.1**	**15.2**	**35.5**	67.5	21.7	52.5	10.1	15.3	36.1	68.9	21.7	53.5	9.5	15.2	35.8	65.6	21.2	**55.0**
Ours *w/o* visual sentinel	9.7	14.5	34.4	63.1	21.0	52.2	9.9	14.7	34.8	65.5	20.8	52.9	9.8	14.8	35.0	64.2	20.9	54.3
Ours	9.9	14.9	35.3	67.3	**22.2**	52.7	**10.8**	**15.7**	**36.4**	**71.3**	**22.0**	**53.9**	**10.9**	**15.8**	**36.2**	**70.4**	**21.8**	**55.0**

Table 4: Controllability via a set of regions, on the test portion of Flickr30K Entities.



A dog holding a frisbee in its mouth.

A dog standing in the grass with a frisbee in its mouth.

A cat laying on a car.

A cat laying on the hood of a car.

A man in a black jacket skiing down a hill.

A man on skis down a snow covered slope.

A train traveling next to a building.

A train traveling down the tracks next to a building.

A zebra standing in the snow.

A zebra standing in the snow next to a wall.

A man and a child flying a kite in a field.

A man and a child flying a kite.

A man hitting a tennis ball.

A man standing on a tennis court with a fence.

A boy wearing a helmet holding a baseball bat.

A boy in a green jersey holding a baseball bat.

A man standing next to a bus on the street.

A bus driving down a street.

A person on a snowboard in the snow.

A person on a snowboard in the snow with a ski lift.

A pizza and a beer on a table.

A pizza on a plate next to a glass of beer.

A dog with a collar standing next to a fire hydrant.

A dog standing in the grass next to a fire hydrant.

Figure 4: Additional sample results of controllability via a set of regions. Different colors show the control set and the associations between chunks and regions.