Adversarial Semantic Alignment for Improved Image Captions
(Supplementary Material)

Pierre Dognin, Igor Melnyk, Youssef Mroueh, Jerret Ross & Tom Sercu
IBM Research, Yorktown Heights, NY

A. Semantic Score

Semantic scores was first introduced int the context of image retrieval where it achieves state of the art performance [18]. Some examples of the properties of semantic scores are given in Table 4.

COCO validation image	Set	Semantic Score	Captions
	Set A	0.181052	female tennis player reaches back to swing at the ball
		0.210224	a woman on a court swinging a racket at a ball
		0.181592	a woman in a gray top is playing tennis
		0.251200	the woman is playing tennis on the court
		0.145646	a woman prepares to hit a tennis ball with a racket
	Set B	0.008990	a clear refrigerator is stocked up with food
		0.005519	a store freezer is shown with food inside
		-0.014052	a refrigerated display case is full of dairy groceries
		0.011076	a close up of a commercial refrigerator with food
		-0.029001	a large cooler with glass doors containing mostly dairy products
	Set C	0.054441	a giraffe reaches back to swing at the ball
		0.123822	female tennis player reaches back to swing at the boat
		0.152860	male tennis player reaches back to swing at the ball
		0.067289	female football player reaches back to swing at the ball
	Set D	0.152860	male tennis player reaches back to swing at the ball
		0.164755	female tennis fan reaches back to swing at the ball
		0.152524	female tennis player looks back to swing at the ball
		0.100098	female flute player reaches back to swing at the ball
	Set E	0.114010	female tennis player swing ball
		0.031566	female player swing ball
		0.084016	tennis player swing ball
		0.115490	tennis player ball
		0.092226	tennis player
		-0.044019	tennis
		-0.001948	ball ball ball ball

Table 4: Semantic scores for various captions given an image from COCO validation set. Set A is composed of the 5 ground truth captions provided by COCO. Semantic scores are in between .14 and .25 for a possible range of [-1,1] being a cosine distance. Set B is made of captions from another randomly selected image in the validation set. The scores are clearly much worse (smaller) when captions do not match the image visual cues. Set C is a one-word modification set of the first caption in Set A. Semantic scores are all lower compared to the original caption. In Set C, we want to see if the metric is solely sensitive the main visual cues and if it can pick up subtle differences like gender. Again, all the scores are still lower, even if closer to the original caption's score. In Set E, we are trying to break the metric by narrowing down to only factual words and objects. The combined knowledge of visual and text correlation penalize simplistic descriptive list of words. This does not imply that the metric cannot be fooled, but it seems resilient to obvious gaming like repeating words of some visual cues.

B. Experimental Results: Complete Tables

We report here CIDEr, BLEU4, ROUGEL, METEOR, semantic scores, and vocabulary coverage for all models mentioned in this work, both COCO and OOC sets. Table 5 presents all GAN results as average (± standard deviation) over 4 models with different random seeds. Table 6 presents all our ensemble results.

Table 5: Collection of results for all models mentioned in this work. We provide commonly used CIDEr, BLEU4, ROUGEL, METEOR scores, as well as semantic scores, and percentage of vocabulary coverage for both COCO and OOC. Results are averaged from 4 models from independent trainings. We report mean and standard deviation for all metrics when available.

	COCO Test Set											
	CIDEr		BLEU4		ROUGEL		METEOR		Semantic Score		Vocabulary Coverage	
CE	101.6	±0.4	0.312	±.001	0.542	±.001	0.260	±.001	0.186	±.001	9.2	±0.1
CIDEr-RL	**116.1**	±0.2	**0.350**	±.003	**0.562**	±.001	0.269	±.000	0.184	±.001	5.1	±0.1
GAN$_1$(SCST, Co-att, $\log(D)$)	97.5	±0.8	0.294	±.002	0.532	±.001	0.256	±.001	0.190	±.000	11.0	±0.1
GAN$_2$(SCST, Co-att, $\log(D)$+5×CIDEr)	111.1	±0.7	0.330	±.004	0.555	±.002	**0.271**	±.002	**0.192**	±.000	7.3	±0.2
GAN$_3$(SCST, Joint-Emb, $\log(D)$)	97.1	±1.2	0.287	±.005	0.530	±.002	0.256	±.002	0.188	±.000	**11.2**	±0.1
GAN$_4$(SCST, Joint-Emb, $\log(D)$+5×CIDEr)	108.2	±4.9	0.325	±.017	0.551	±.008	0.267	±.004	0.190	±.000	8.3	±1.6
GAN$_5$(Gumbel Soft, Co-att, $\log(D)$)	93.6	±3.3	0.282	±.015	0.524	±.007	0.253	±.007	0.187	±.002	11.1	±1.2
GAN$_6$(Gumbel ST, Co-att, $\log(D)$)	95.4	±1.5	0.298	±.009	0.531	±.005	0.249	±.004	0.184	±.003	10.1	±0.9
GAN$_7$(Gumbel ST, Co-att, $\log(D)$+FM)	92.1	±5.4	0.289	±.020	0.523	±.015	0.243	±.011	0.175	±.006	8.6	±0.8
G-GAN [4] from Table 1	79.5		0.207		0.475		0.224		–		–	
CE* – * for non-attentional models	87.6	±1.2	0.275	±.003	0.516	±.003	0.242	±.001	0.175	±.002	9.9	±0.8
CIDEr-RL*	100.4	±7.9	0.305	±.018	0.536	±.010	0.253	±.006	0.173	±.002	6.8	±1.4
GAN$_1$*(SCST, Co-att, $\log(D)$)	89.7	±0.9	0.276	±.000	0.518	±.001	0.246	±.001	0.184	±.001	13.2	±0.2
GAN$_2$*(SCST, Co-att, $\log(D)+5$×CIDEr)	103.1	±0.5	0.311	±.003	0.542	±.001	0.261	±.001	0.183	±.001	7.1	±0.2
GAN$_3$*(SCST, Joint-Emb, $\log(D)$)	90.7	±0.1	0.277	±.002	0.520	±.000	0.248	±.001	0.181	±.001	12.9	±0.1
GAN$_4$*(SCST, Joint-Emb, $\log(D)+5$×CIDEr)	102.7	±0.4	0.315	±.000	0.542	±.000	0.260	±.001	0.182	±.001	7.7	±0.1
	OOC (Out of Context)											
	CIDEr		BLEU4		ROUGEL		METEOR		Semantic Score		Vocabulary Coverage	
CE	42.2	±0.6	0.168	±.005	0.413	±.003	0.169	±.001	0.118	±.001	2.8	±0.1
CIDEr-RL	45.0	±0.6	0.177	±.002	0.417	±.004	0.170	±.003	0.117	±.002	2.1	±0.0
GAN$_1$(SCST, Co-att, $\log(D)$)	41.0	±1.6	0.161	±.013	0.406	±.006	0.168	±.003	**0.124**	±.000	3.2	±0.1
GAN$_2$(SCST, Co-att, $\log(D)+5$×CIDEr)	**45.8**	±0.9	0.179	±.014	0.417	±.005	**0.173**	±.001	0.122	±.002	2.8	±0.1
GAN$_3$(SCST, Joint-Emb, $\log(D)$)	41.8	±1.6	0.162	±.006	0.404	±.006	0.167	±.002	0.122	±.001	**3.3**	±0.0
GAN$_4$(SCST, Joint-Emb, $\log(D)+5$×CIDEr)	45.4	±1.4	**0.180**	±.011	**0.418**	±.005	**0.173**	±.002	0.122	±.003	2.8	±0.2
GAN$_5$(gumbel soft, Co-att, $\log(D)$)	38.3	±3.7	0.154	±.020	0.406	±.006	0.164	±.006	0.121	±.004	**3.3**	±0.3
GAN$_6$(gumbel-ST, Co-att, $\log(D)$)	38.5	±1.9	0.148	±.005	0.407	±.004	0.161	±.005	0.116	±.004	3.0	±0.2
GAN$_7$(gumbel-ST, Co-att, $\log(D)$+FM)	36.8	±2.3	0.154	±.012	0.396	±.009	0.157	±.006	0.110	±.005	2.5	±0.2
CE*	32.0	±0.4	0.132	±.007	0.392	±.002	0.152	±.002	0.103	±.002	2.6	±.1
CIDEr-RL*	33.4	±1.4	0.145	±.009	0.394	±.006	0.154	±.003	0.101	±.003	2.1	±.2
GAN$_1$*(SCST, Co-att, $\log(D)$)	30.8	±1.0	0.127	±.001	0.383	±.006	0.155	±.003	0.111	±.001	3.4	±0.1
GAN$_2$*(SCST, Co-att, $\log(D)+5$×CIDEr)	33.7	±1.9	0.145	±.011	0.391	±.004	0.157	±.001	0.108	±.001	2.7	±0.1
GAN$_3$*(SCST, Joint-Emb, $\log(D)$)	30.8	±2.1	0.126	±.009	0.380	±.004	0.153	±.002	0.108	±.001	3.5	±0.1
GAN$_4$*(SCST, Joint-Emb, $\log(D)+5$×CIDEr)	33.3	±2.4	0.144	±.016	0.391	±.006	0.157	±.004	0.106	±.000	2.7	±0.1

C. Semantic and Discriminator Scores Correlation over Training Epochs

We are interested in the correlation between the semantic scores and discriminator scores of image captions as well as its evolution along the process of SCST GAN training. We provide scatter plots for the Joint-Embedding discriminator [4] across training in Figure 9. This GAN model was trained over 40 epochs with a discriminator pretrained on 15 epochs of data.

We compare semantic scores and discriminator scores over training epochs given the ground truth (GT) caption for each image in the COCO Test set (5K images). Each GT caption being fixed, we can observe the evolution of the semantic and discriminator score without any other effects. Figure 9 show the semantic score, discriminator score pairs for each image (one

Table 6: Collection of ensembling results for GAN models from Table 2. We provide commonly used CIDEr, BLEU4, ROUGEL, METEOR scores, as well as semantic scores, and percentage of vocabulary coverage for both COCO and OOC.

		COCO Test Set					
		CIDEr	BLEU4	ROUGEL	METEOR	Semantic Score	Vocabulary Coverage
(CE and RL Baselines)	Ens_{CE}(CE)	105.8	0.327	0.553	0.266	0.189	8.4
	Ens_{RL}(CIDEr-RL)	**118.9**	**0.359**	**0.568**	0.273	0.186	5.0
(SCST, Co-att, $*$)	$\text{Ens}_1(\text{GAN}_1)$	102.6	0.314	0.543	0.262	**0.195**	9.9
	$\text{Ens}_2(\text{GAN}_2)$	115.1	0.347	0.566	**0.277**	0.194	7.0
	$\text{Ens}_{12}(\text{GAN}_1,\text{GAN}_2)$	113.2	0.344	0.564	0.274	**0.195**	7.3
(SCST, Joint-Emb, $*$)	$\text{Ens}_3(\text{GAN}_3)$	109.8	0.331	0.556	0.270	0.193	8.5
	$\text{Ens}_4(\text{GAN}_4)$	113.0	0.343	0.562	0.274	0.193	7.6
	$\text{Ens}_{34}(\text{GAN}_3,\text{GAN}_4)$	111.1	0.335	0.558	0.271	0.193	8.1
(Gumbel $*$, Co-att, $*$)	$\text{Ens}_5(\text{GAN}_5)$	100.1	0.307	0.538	0.259	0.191	**10.0**
	$\text{Ens}_6(\text{GAN}_6)$	99.6	0.313	0.541	0.253	0.187	9.3
	$\text{Ens}_7(\text{GAN}_7)$	100.2	0.321	0.543	0.254	0.180	7.8
	$\text{Ens}_{567}(\text{GAN}_5,\text{GAN}_6,\text{GAN}_7)$	103.2	0.327	0.550	0.258	0.188	8.7
(SCST+Gumbel Soft, Co-att, $*$)	$\text{Ens}_{125}(\text{GAN}_1,\text{GAN}_2,\text{GAN}_5)$	112.4	0.343	0.562	0.273	**0.195**	7.7
		OOC (Out of Context					
		CIDEr	BLEU4	ROUGEL	METEOR	Semantic Score	Vocabulary Coverage
(CE and RL Baselines)	Ens_{CE}(CE)	44.8	0.177	0.423	0.172	0.122	2.6
	Ens_{RL}(RL)	48.8	**0.198**	0.427	0.175	0.122	2.1
(SCST, Co-att, $*$)	$\text{Ens}_1(\text{GAN}_1)$	44.8	0.175	0.422	0.172	**0.129**	**3.0**
	$\text{Ens}_2(\text{GAN}_2)$	48.3	0.189	0.429	0.176	0.127	2.7
	$\text{Ens}_{12}(\text{GAN}_1+4\times\text{GAN}_2)$	49.9	0.197	**0.437**	0.178	**0.129**	2.6
(SCST, Joint-Emb, $*$)	$\text{Ens}_3(\text{GAN}_3)$	48.5	**0.198**	0.429	0.175	0.127	2.8
	$\text{Ens}_4(\text{GAN}_4)$	48.0	0.185	0.432	0.178	0.127	2.7
	$\text{Ens}_{34}(\text{GAN}_3+4\times\text{GAN}_4)$	**50.1**	0.195	0.435	0.177	0.127	2.8
(Gumbel $*$, Co-att, $*$)	$\text{Ens}_5(\text{GAN}_5)$	43.1	0.169	0.420	0.170	0.127	**3.0**
	$\text{Ens}_6(\text{GAN}_6)$	41.0	0.155	0.420	0.165	0.122	2.8
	$\text{Ens}_7(\text{GAN}_7)$	38.9	0.166	0.413	0.164	0.113	2.3
	$\text{Ens}_{567}(\text{GAN}_5,\text{GAN}_6,\text{GAN}_7)$	41.8	0.167	0.418	0.164	0.121	2.7
(SCST+Gumbel Soft, Co-att, $*$)	$\text{Ens}_{125}(\text{GAN}_1,\text{GAN}_2,\text{GAN}_5)$	49.8	**0.198**	0.436	**0.179**	**0.129**	2.7

point per image) for the joint embedding discriminator. Since the GT captions are fixed, the semantic scores will be identical across epochs. From the first epoch, the joint embedding discriminator provides a wide range of scores with most scores close to the 0.0 and 1.0 min/max values. Quickly the points cluster into a 'sail' like shape in the lower right corner, away from the min/max edges. The color assigned to each point is directly linked to the semantic scores assigned at the first epoch of training. You can therefore have a small visual cue of the movement of these points from epoch to epoch and witness the discriminator learning how to distinguish real and fake captions.

D. Human Evaluation

In this section we present the details of our evaluation protocol for our captioning models on Amazon MTurk. All images are presented to 5 workers and aggregated in mean opinion score (MOS) or majority vote.

Turing Test. In this setting we give human evaluators an image with a sentence either generated from our GAN captioning models or the ground truth. We ask them whether the sentence is human generated or machine generated. Exact instructions are: "Is this image caption written by a human? Yes/No. The caption could be written by a human or by a computer, more or less 50-50 chance."

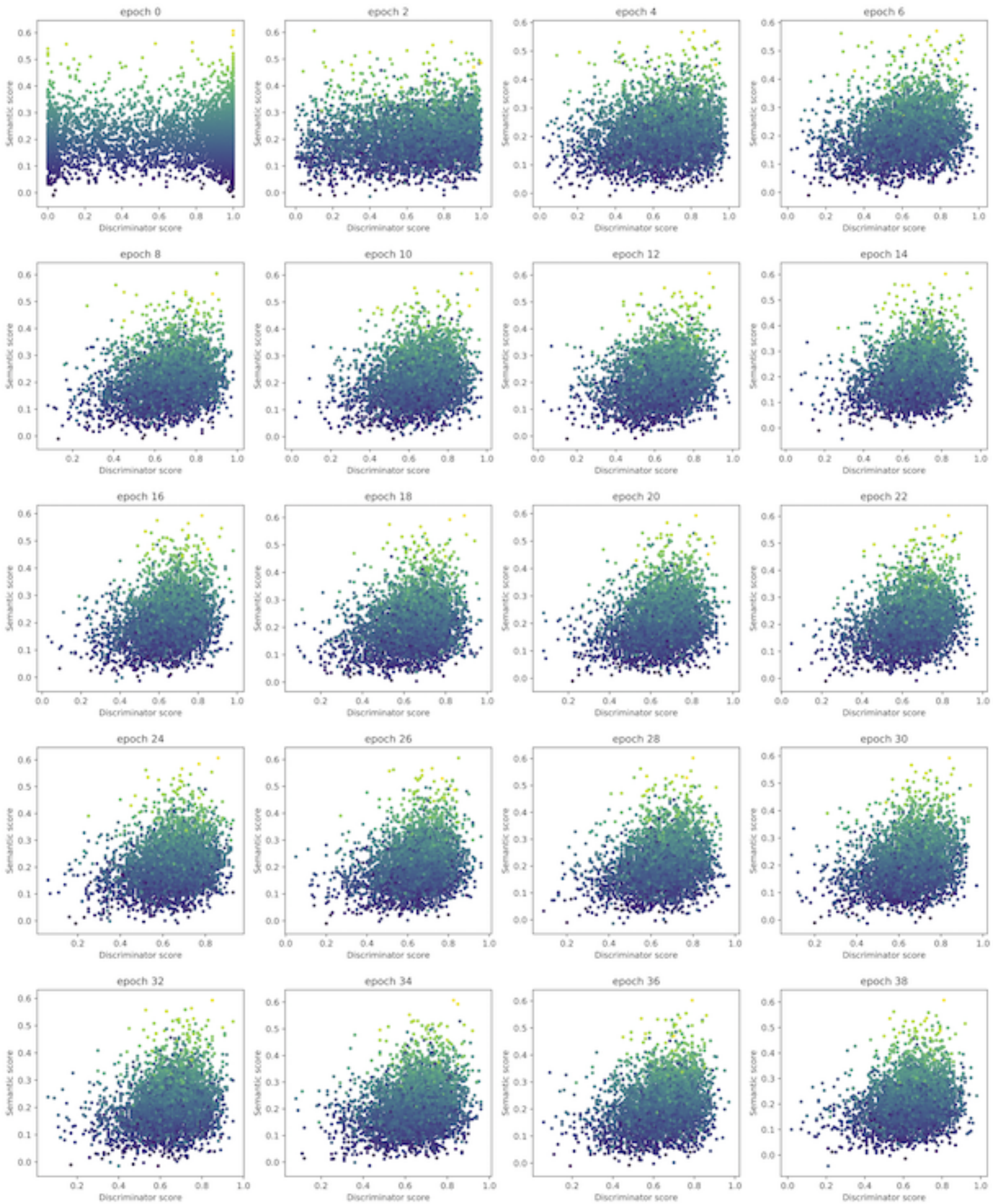


Figure 9: Semantic vs. Discriminator scores across 40 training epochs for ground truth captions using the joint embedding discriminator [4].

Fine Grained Evaluation and Model Comparison. In this experiment we give human evaluators an image and a set of 3 captions: Generated by CE trained model, SCST CIDEr trained model, and a GAN model. We ask them to rate each sentence on a scale of one to five. After rating, the worker chooses the caption he/she thinks is best at describing the image. In Section 4, we provide results for Mean Opinion Score and Majority vote based of this interface (see Figure 10) and Table 7.

E. Experimental Protocol SCST vs. Gumbel

In Figure 11, we show that all our Gumbel Methods trained effectively. We plot the Discriminator scores (averaged over minibatch) during training with the 3 reported Gumbel models. Generated sentences get roughly 0.5, random sentences around 0.1, real sentences around 0.75. Hence, the Discriminator can correctly distinguish real from random, and generated sentences.

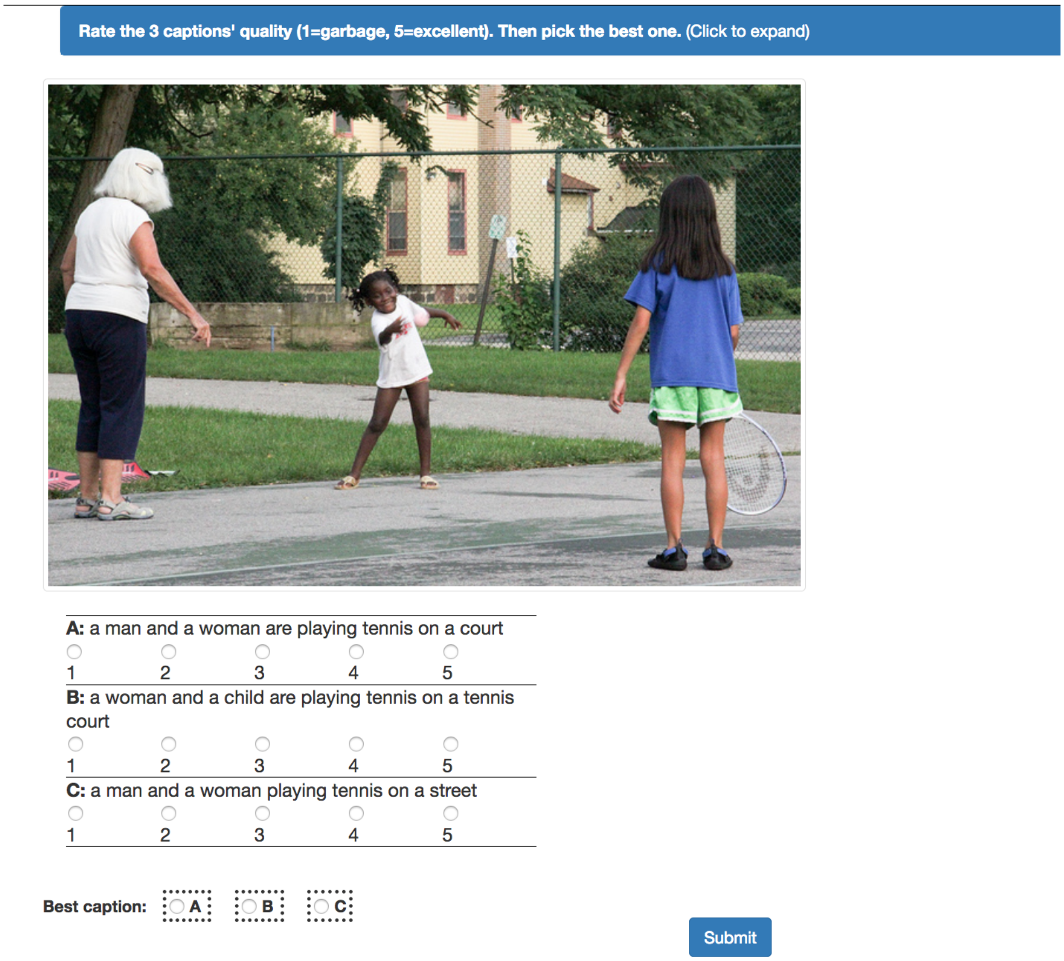


Figure 10: The interface of "Fine Grained Evaluation".

Table 7: MOS and semantic scores collected from Amazon MTurk.

	COCO Test		OOC	
	Semantic Score	MOS	Semantic Score	MOS
Ens_{CE}(CE)	0.189	3.222	0.122	3.065
Ens_{RL}(CIDEr-RL)	0.186	3.297	0.122	3.097
Ens_1(SCST, Co-att, $\log(D)$)	**0.195**	3.398	–	–
Ens_2(SCST, Co-att, $\log(D) + 5 \times \text{CIDEr}$)	0.194	**3.442**	0.127	3.107
Ens_3(SCST, Joint-Emb, $\log(D)$)	0.193	3.286	–	–
Ens_5(Gumbel Soft, Co-Att,$\log(D)$)	0.191	3.138	–	–
Ens_7(Gumbel ST, Co-Att, $\log(D) + \text{FM}$)	0.180	3.235	–	–

This indicates a healthy training of all Gumbel Methods.

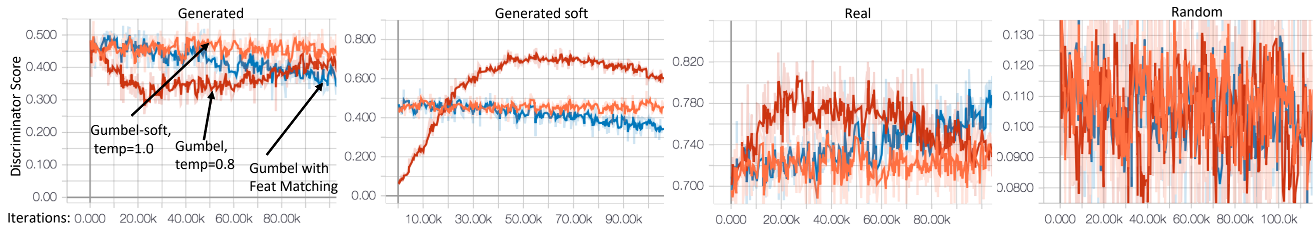
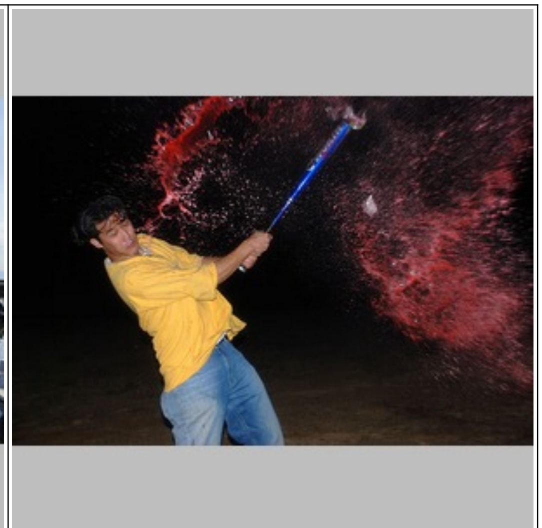


Figure 11: Discriminator scores across different training Gumbel Methods.

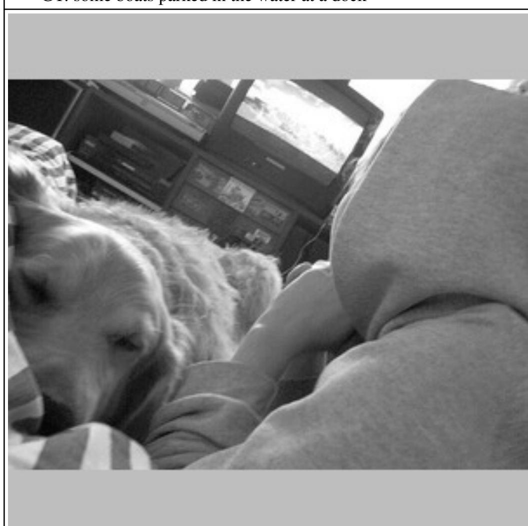
F. Examples of Generated Captions

In this section we present several examples of captions generated from our model. In particular, Figure 12 and Figure 13 show captions for randomly picked images (from COCO and OOC respectively) which provide a good description of the image content. We do the opposite in Figure 14 and Figure 15 where examples of bad captions are provided for COCO and OOC respectively.

GAN: a group of boats are docked in a harbor
CE: a group of boats sitting in the water
RL: a group of boats sitting in the water
GT: some boats parked in the water at a dock



GAN: a man holding a baseball bat at night
CE: a man is holding a bat in a dark
RL: a man holding a baseball bat at a ball
GT: a boy in yellow shirt swinging a baseball bat



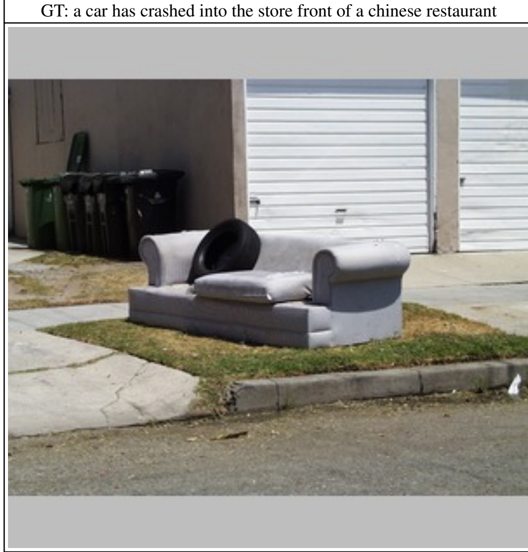
GAN: a dog laying down on a person 's lap
CE: a dog is sleeping on a couch with a person
RL: a dog laying on a couch with a person
GT: a dog that is laying next to another person

GAN: a bunch of umbrellas hanging from a wall in a store
CE: a bunch of umbrellas that are hanging from a wall
RL: a group of umbrellas hanging from a store
GT: a bunch of umbrellas that are behind a glass

Figure 12: Cherry-picked examples on the COCO validation set.

GAN: a store front with a car parked in front of it
CE: a store with a sign on the side of it
RL: a building with a sign on the side of it
GT: a car has crashed into the store front of a chinese restaurant

GAN: a bed sitting in the middle of a forest
CE: a bed with a green blanket on it
RL: a bed in a forest with a table
GT: a bed lies on top of a clover field in a forest



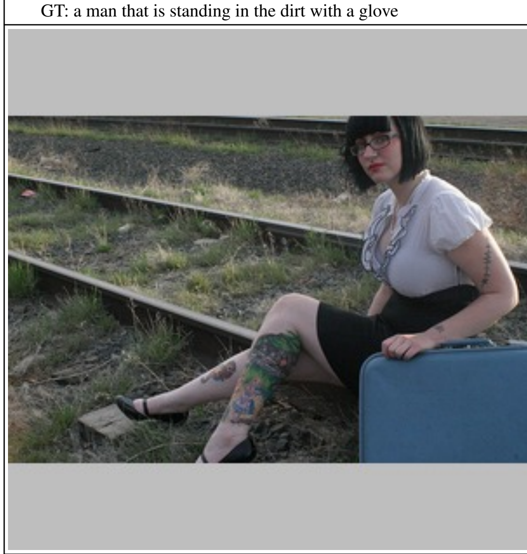
GAN: a couch sitting in front of a house with a trash can
CE: a white couch sitting in front of a house
RL: a couch sitting in front of a house
GT: a white couch on top of a grass curb with a black table in the background

GAN: a large passenger jet taking off from a busy street
CE: a large passenger jet sitting on top of a runway
RL: a group of cars parked on the runway at an airplane
GT: an airplane descends very close to traffic stuck at a red light

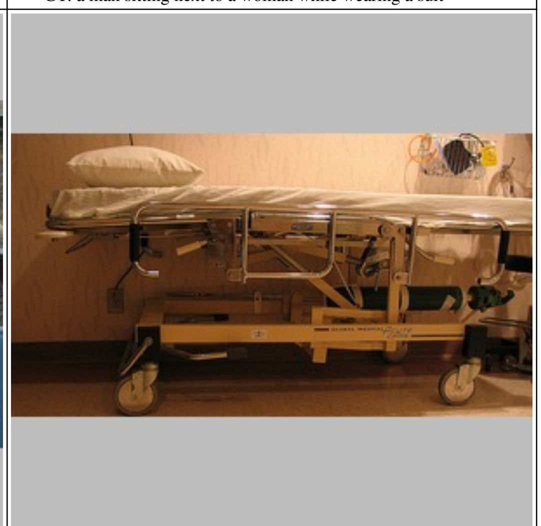
Figure 13: Cherry-picked examples on the Out of Context (OOC) set.

GAN: a baseball player swinging a bat at a ball
CE: a baseball player is swinging a bat at a ball
RL: a baseball player swinging a bat at a ball
GT: a man that is standing in the dirt with a glove

GAN: a black and white photo of two men in suits
CE: a man and a woman standing next to each other
RL: a black and white photo of a man and a woman
GT: a man sitting next to a woman while wearing a suit



GAN: a woman sitting on a bench looking at her cell phone
CE: a woman sitting on a bench in a park
RL: a woman sitting on the ground next to a bench
GT: a woman is sitting with a suitcase on some train tracks



GAN: a bike that is in a room with a bike
CE: a bicycle with a bicycle and a bicycle in a room
RL: a room with a bed and a table in a room
GT: the hospital bed is metal and has wheels

Figure 14: Lime-picked examples on the COCO test set.

	
GAN: a man standing on the side of the road with a skateboard CE: a man standing on the side of a road with a cell phone RL: a man standing on the side of a road with a cell phone GT: a woman holds a drink can while holding the door to a refrigerator that is sitting on the asphalt of a street	GAN: two people walking on a beach with a dog CE: a couple of people walking on a beach with a dog RL: a group of people walking on a beach with a dog GT: a lady is flying a chair as if its a kite while walking along the water edge
	
GAN: a couple of chairs and a blue beach chairs on a beach CE: a couple of chairs sitting next to each other on a beach RL: a group of chairs and a table in the beach GT: a picture of a chair on an empty beach with a laptop on the arm	GAN: a painting of a vase in front of a fire hydrant CE: a painting of a fire place in a room RL: a bedroom with a bed and a clock on the wall GT: a white goat wearing a gold crown sits on a gold bed

Figure 15: Lime-picked examples on the Out of Context (OOC) set.