Describing like Humans: on Diversity in Image Captioning —*Supplementary Material*

Qingzhong Wang and Antoni B. Chan
Department of Computer Science
City University of Hong Kong
qingzwang2-c@my.cityu.edu.hk, abchan@cityu.edu.hk

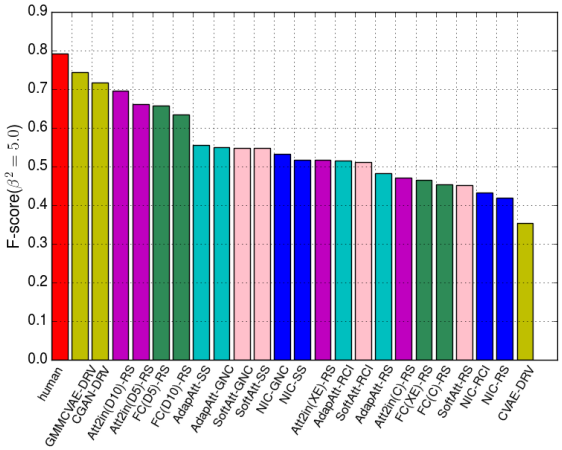

Figure 1. The F-scores (using Self-CIDEr) of different models.

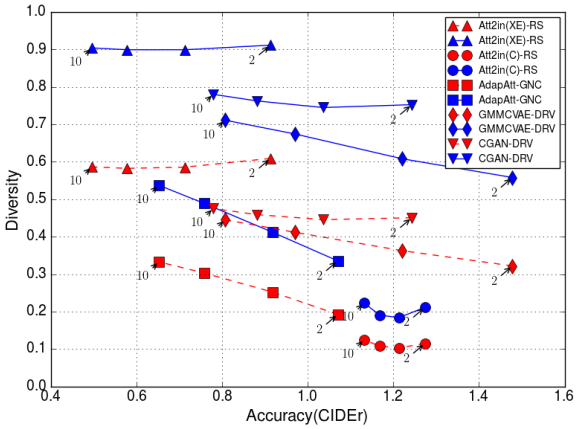

Figure 2. The effects of different numbers of captions on diversity. We use $m \in \{2, 5, 8, 10\}$. The dash lines represent LSA-based diversity and the solid lines represent Self-CIDEr diversity.

1. F-measure

The accuracy and diversity tradeoff is summarized using the F-measure, $F = \frac{(1+\beta^2)div \cdot acc}{\beta^2 div + acc}$. $\beta > 1$ will weight accuracy more, while $1 > \beta \geq 0$ will weight diversity more.

Figure 1 shows the F-scores that takes both diversity and accuracy into account. In this paper, we use $\beta^2 = 5.0$, which considers accuracy is more important than diversity. The reason for using a larger β is that diverse captions that do not describe the image well (low accuracy) could be meaningless. The F-score of human annotations is the highest, and much higher than the models that generate only a single accurate caption for each image. CGAN and GMMCVAE that are specifically designed to generate diverse captions also obtain high F-scores, which are closer to human performance, and this is consistent with Figure 4 in our paper. For Att2in and FC, applying retrieval reward is a better way to achieve both diverse and accurate captions. Looking at the models using RS, more advanced models obtain higher F-score, which is also consistent with Figure 4 in the paper. Note that the scales of diversity and accuracy scores are different.

2. The effects of number of captions

We next consider how many captions should be generated to evaluate diversity. Here, we use Att2in(XE)-RS, Att2in(C)-RS and AdapAtt-GNC, which are exemplar methods located in different areas in the DA plot (Figure 4 in the paper), and GMMCVAE and CGAN, since they are the models with highest F-scores. We first rank the captions for each image based on accuracy, then we select the top 2, 5, 8 and 10 captions to calculate the diversity scores (see Figure 2).

For Att2in(XE)-RS, which obtains very high diversity but low accuracy, the number of captions has small effect on the diversity. This is also seen in Att2in(C)-RS, which obtains low diversity but higher accuracy. The reason is that the captions generated by Att2in(XE)-RS are nearly always completely different with each other, while captions generated by Att2in(C)-RS are almost always the same. Therefore, increasing the number of generated captions does not affect the diversity for these models. For models that well

balance diversity and accuracy (e.g., CGAN and GMMC-VAE), more captions leads to higher diversity, and to some extent, diversity is linearly proportional to the number of captions. Therefore, we suggest that if a model is able to generate diverse captions, more captions should be generated to evaluate its diversity. Although the number of captions has an effect on diversity scores, a better model generally obtains higher diversity.

3. Comparison with mBLEU

Given a set of captions $\mathcal{C} = \{c_1, c_2, \cdots, c_m\}$, mBLEU [**?**] is computed as follows:

$$mBLEU_n = \frac{1}{m}\sum_{i=1}^{m} BLEU_n(c_i, \mathcal{C}\backslash i), \qquad (1)$$

where n represents using n-gram, $BLEU_n()$ represents the BLEU function and $\mathcal{C}\backslash i$ denotes the set of captions without the i-th caption. A higher mBLEU score indicates lower diversity, here we use $1 - mBLEU_n$ to measure the diversity of $\mathcal{C}$, thus, a higher score indicates higher diversity. We also consider the mixed mBLEU score, which is the weighted sum of $mBLEU_n$, *i.e.*, $\sum_{n=1}^{4} \omega_n mBLEU_n$, and in this paper we set $\omega_n = \frac{1}{4}$.

Figure 3 shows the correlation between our Self-CIDEr diversity metric and the mBLEU diversity metric. Figure 4 shows the rankings of different models based on Self-CIDEr diversity and mBLEU-mix diversity.

Figure 5 shows the captions generated by NIC-SS and FC(D10)-RS and the corresponding diversity scores, and figure 6 shows the sets of captions generated by human, SoftAtt-RS and AdapAtt-RS and the corresponding diversity scores.

Another advantage of using LSA-based and Self-CIDEr diversity metrics is that we can project the captions into a latent semantic space vis decomposition, thus we are able to visualize each set of captions in the space composed of 2 latent semantics (see section 4).

4. Visualizing Captions via Decomposition

In this section we visualize the captions using LSA and our proposed kernelized method (Self-CIDEr) to project captions into the semantic space, thus we can see what the captions are talking about. Given a set of captions $\mathcal{C}$, we first construct a dictionary $\mathcal{D}$ and then each caption is represented by bag-of-word features, and $\mathcal{C}$ is represented by the "word-caption" matrix $\mathbf{M}$ (we have described the details in our paper). To better visualize the captions, we use stop words in LSA[1].

Recall that $\mathbf{M} = \mathbf{USV}^T$ using SVD and we select the 5 largest singular values and their corresponding row vectors $[\mathbf{v}_1; \mathbf{v}_2; \cdots; \mathbf{v}_5]$ in $\mathbf{V}^T$, where $\mathbf{v}_i \in \mathbb{R}^m$, m denotes the number of captions. In LSA, $\mathbf{v}_i$ reflects the relationship between captions and the i-th latent semantic. Hence, the j-th caption c_j can be represented by a 5-D vector $[v_1^j, v_2^j, \cdots, v_5^j]$, where v_i^j is the j-th elements of $\mathbf{v}_i$. Similarly, for the kernelized method, we can decompose the kernel matrix $\mathbf{K}$, thus $\mathbf{K} = \mathbf{V\Lambda V}^T$ and the 5-D vector $[v_1^j, v_2^j, \cdots, v_5^j]$ can also represent a caption in the semantic space. In this paper we use radar charts to illustrate the correlations between captions and latent semantics. A positive value could represent that a caption contains this semantic, a negative or null value could indicate that a caption does not contain this semantic and it describes another thing. Figure 7 to 9 show the captions generated by different models and each caption in the 5-D latent semantic space. Note that the semantics in LSA and Self-CIDEr could be different.

5. Captions Generated by RL-based Methods

We show some generated captions of RL-based methods with different combinations of loss functions (see FigureS 10 and 11).

[1] https://www.nltk.org/book/ch02.html

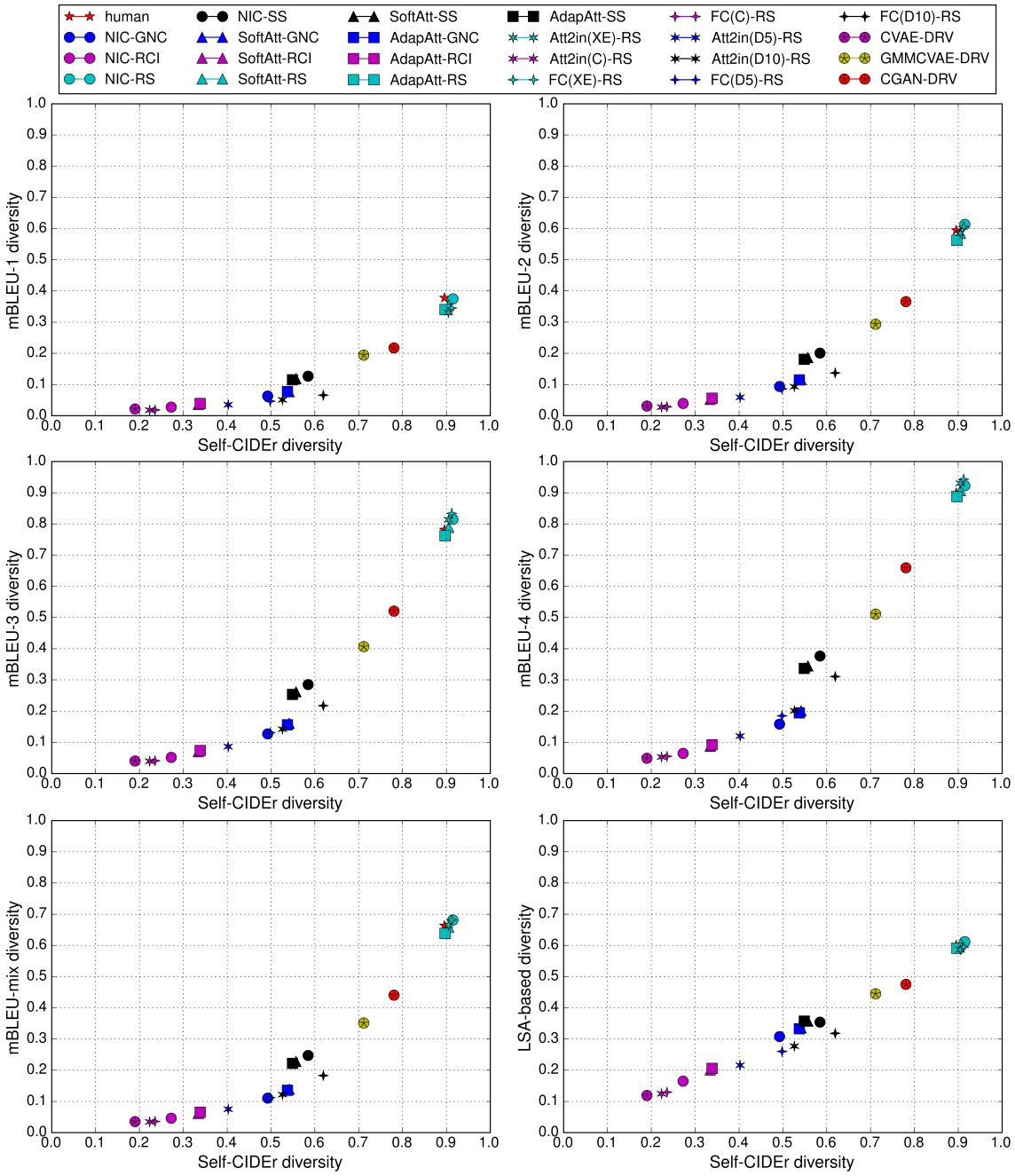

Figure 3. The correlation between our Self-CIDEr diversity and mBLEU scores. We use 5 mBLEU scores—uni-, bi-, tri-, quad- and mixed BLEU scores. To some extent, our Self-CIDEr diversity score is consistent with mBLEU diversity scores, and it shows an exponential correlation between mBLEU diversity scores and Self-CIDEr diversity score. The reason is that in Self-CIDEr we use $-\log_m(r)$ as the final diversity score. The biggest difference is that our Self-CIDEr assigns a higher diversity score to FC(D10)-RS, by contrast the mBLEU metric assigns a higher score to NIC-SS (see the ranking of different models in Figure 4). Recall that SS approach applies synonyms to replace the words in a captions, which just changes the words but could not change the semantics, and Self-CIDEr diversity metric that is derived from latent semantic analysis (LSA) pays much attention to *semantic diversity*, hence, using synonyms could result in low Self-CIDEr diversity. Moreover, mBLEU-1,2,3,4 and mix could assign differnt rankings to the same model, *e.g.*, FC(D5)-RS is ranked below NIC-GNC using mBLEU-1,2, whereas mBLEU-4 assigns a higher score to FC(D5)-RS than NIC-GNC. In contrast, both of them obtain similar diversity score using mBLEU-3,mix and Self-CIDEr. The correction between LSA-based and Self-CIDEr is roughly linear.

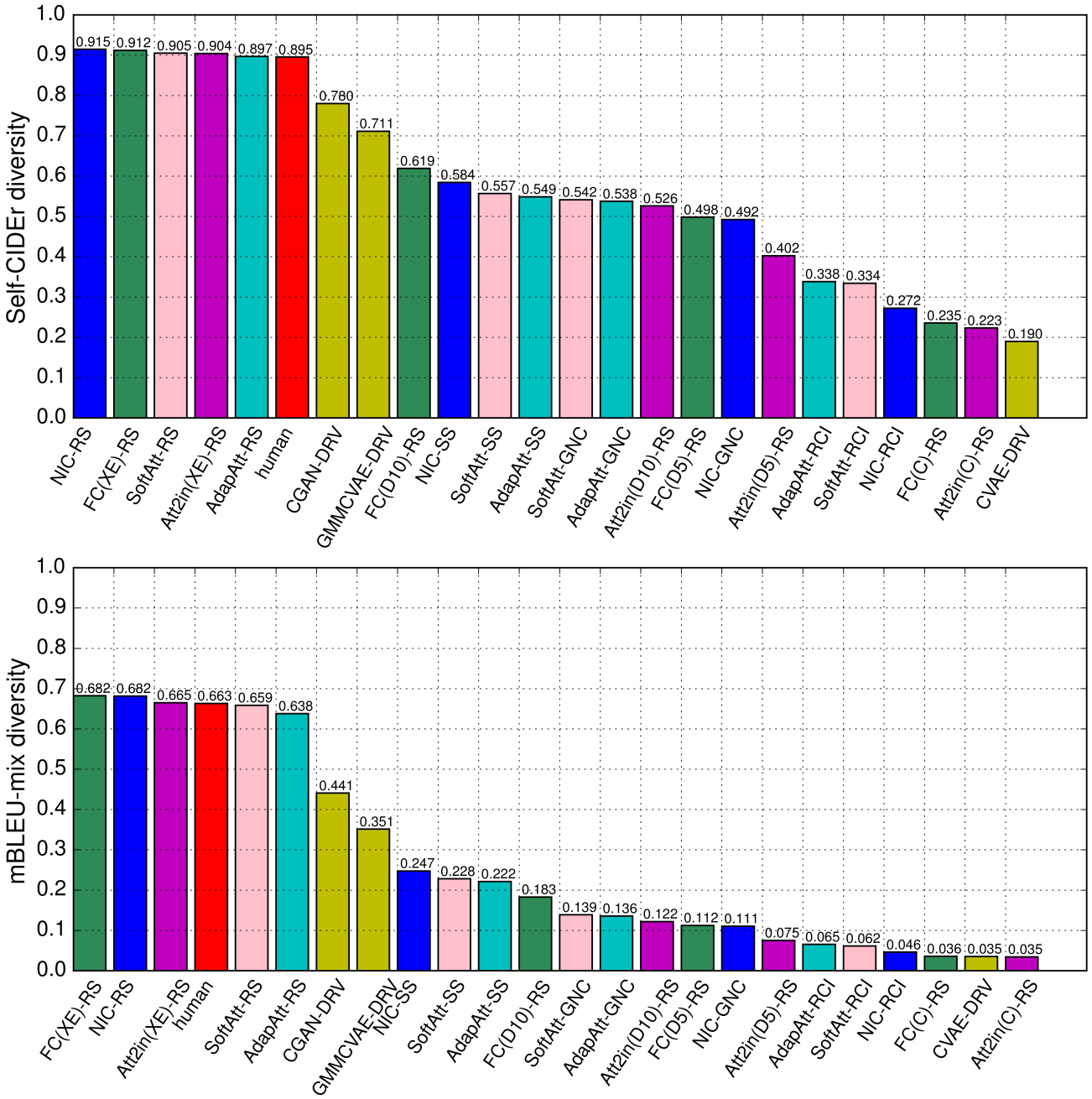

Figure 4. Model ranking by Self-CIDEr diversity and the mixed mBLEU diversity. Generally, mBLEU-mix diversity scores are lower than Self-CIDEr diversity scores, although both of them account for uni-, bi-, tri- and quad-grams. In most cases, the two diversity metrics provide consistent ranking, except for the rankings of human and FC(D10)-RS. Human is ranked below SoftAtt-RS and AdapAtt-RS using Self-CIDEr , while mBLEU-mix ranks human above them. Another difference is that FC(D10)-RS is treated as a more diverse model than the SS models using Self-CIDEr diversity metric, in contrast, using mBLEU-mix, FC(D10)-RS obtains a lower diversity score.

1. a train traveling through tracks tending to a loading dock
2. a train traveling down tracks next to a loading subway
3. a train traveling down tracks next toa a loading platform
4. a train moving down tracks next toa a loading platform
5. a train traveling sown tacks next amtrak a loading platform
6. a train traveling down tracks next to a loading platform
7. a train traveling down tracks next tot a loading platform
8. a train traveling down tracks next to a loading platform
9. a train traveling across tracks next amtrak a loads platform
10. a train moving down runway next to a loading station

NIC-SS Self-CIDEr: 0.694
mBLEU-1: 0.178 mBLEU-2: 0.282
mBLEU-3: 0.456 mBLEU-4: 0.541

1. a view of a train window in a airport terminal terminal
2. a view of a train overpass in a station overpass bridge
3. a train car driving down a highway overpass overpass gate
4. a car driving down a train station a terminal gate station
5. a view of a train terminal in a terminal terminal gate overpass
6. a view of a train window in a station station terminal
7. a bridge of a train station in a station terminal terminal
8. a train crossing over a train station the gate overpass overpass
9. a view of a train cars in a station gate overpass station
10. a view of a train cars in a terminal terminal overpass

FC(D10)-RS Self-CIDEr: 0.775
mBLEU-1: 0.072 mBLEU-2: 0.189
mBLEU-3: 0.321 mBLEU-4: 0.497

1. a village bus includes standing across the top of the road
2. a city buses is parked on the side of the road
3. a city buses is standing at the side of the street
4. a city bus is parked on the side of the road
5. a suburban bus is parked on the side of the road
6. a city bus is parked on the side of the road
7. a city bus are stopped on the side of the road
8. a city bus is parked on the side of the road
9. a residential bus are pared on the edge of the roadway
10. a city bus is parked on the side of the road

NIC-SS Self-CIDEr: 0.664
mBLEU-1: 0.175 mBLEU-2: 0.261
mBLEU-3: 0.393 mBLEU-4: 0.464

1. a white bus parked in the parking lot doors
2. a bus white bus parked in a parking lot
3. a white bus parked bus parked on the parking lot
4. a bus white bus parked in the parking lot
5. a white bus parked with a bus parking lot
6. a bus white bus parked in a parking lot
7. a white bus parked on a sidewalk bus doors
8. a bus white bus parked parked on the curb
9. a bus white bus parked in a parking lot
10. a passenger bus bus parked on a parking lot

FC(D10)-RS Self-CIDEr: 0.546
mBLEU-1: 0.044 mBLEU-2: 0.113
mBLEU-3: 0.174 mBLEU-4: 0.258

Figure 5. Generated captions of NIC-SS and FC(D10)-RS models and the corresponding diversity scores, where a higher score indicates diverse captions. For the first image (top), Self-CIDEr and mBLEU metrics provide different rankings of NIC-SS and FC(D10)-RS. NIC-SS is ranked below FC(D10)-RS based on Self-CIDEr, while the mBLEU diversity scores of NIC-SS are higher. Looking at the captions, NIC-SS just switches "traveling" to "moving", "down" to "through" or "across" and "platform" to "station", while FC(D10)-RS describes different concepts, such as "train", "car(s)", "airport terminal", "stations", "overpass" and "bridge". Therefore, FC(D10)-RS obtains higher Self-CIDEr diversity score but lower mBLEU diversity scores. For the second image (bottom), Self-CIDEr and mBLEU diversity metrics provide consistent ranking, because both NIC-SS and FC(D10)-RS describe the same thing—*bus parked on*. Comparing the two images, the first one contains more concepts than the second one, and using SS just changes the words but does not change the semantics, in contrast, FC(D10)-RS introduces different concepts to different captions, which could result in different semantics, however, FC(D10)-RS could generate non-fluent sentences.

Human	SoftAtt-RS	AdapAtt-RS
1.white and orange flowers in a glass vase 2.red and white flowers in a vase on a table 3.the flowers are in the vase on display 4.a vase filled with red flowers on a wooden table 5.this vase is holding a bunch of beautiful blooms	1.there is a vase with red and orange flowers in it 2.a vase with a red flower vase on top of a table 3.a vase with a flower in it is sitting on a table top 4.a vase with red and orange leaves in the middle is idly 5.a vase with different types of flowers on a table 6.a vase is holding a set of red leaves and red flowers 7.two red and yellow flowers are on a table 8.a vase with orange coca red flower falling the color of a flower on table 9.a big vase of red floral sits on a window sill 10.a vase with a branch window and branches in the house"	1.there is a clear vase on the dinning table 2.a vase filled with water with various flowers 3.a vase with a red flowering on it hanging 4.both both red are holding different long red and red flowers in each ways 5.a clear vase is holding some little flowers coming from a vase 6.a vase that has an arrangement of red in the ground 7.there is a vase with many red flowers sitting in a corner with small blooms 8.a vase with many very pretty flowers sitting in a green vase 9.a clear glass vase holding a bunch of red berries 10.a vase with orange flowers sitting in a rainy light
Human Self-CIDEr: 0.835 **mBLEU-1: 0.366 mBLEU-2: 0.626** **mBLEU-3: 0.852 mBLEU-4: 1.000**	**SoftAtt-RS Self-CIDEr: 0.843** **mBLEU-1: 0.216 mBLEU-2: 0.433** **mBLEU-3: 0.640 mBLEU-4: 0.798**	**AdapAtt-RS Self-CIDEr: 0.878** **mBLEU-1: 0.358 mBLEU-2: 0.543** **mBLEU-3: 0.731 mBLEU-4: 0.876**

Human	SoftAtt-RS	AdapAtt-RS
1.a zoo keeper on a scale holding a giraffe with a me gusta face 2.the man is carrying a young giraffe in his arms 3.a photo of a man holding a giraffe to find out how much it weighs 4.someone put a face over a baby giraffe that a man is trying to weigh 5.a man lifts a giraffe which seems to have been altered	1.a man is petting the giraffe a neck of frisbees 2.there is a little girl posing with a calf at an exhibit 3.a girl in a backpack looking at a giraffe 4.the person is holding their dog outside near the bike 5.a young zebra is looking at a giraffe 6.a teenage boy posing with a young girl in an outdoor plaza 7.a small child is holding a bat over the gate 8.a person on a skateboard holding a cat 9.an adult giraffe with sticks out of her head 10.the child is feeding the giraffes up on the fence	1.a wild giraffe squatted aside gigantic enclosure one giraffe looks away 2.a young man stands his head as he walks toward a giraffe in a pen 3.a little boy standing over fence to pet a giraffe 4.an adult giraffe is licking a womans hand at the zoo 5.a giraffe standing up against a gate with a ball in his hand 6.there is a little boy that is trying to feed a giraffe 7.a person mate for a giraffe in front of a crowd 8.a giraffe at the zoo reaches into a fence 9.a young man witting on the side of a building while in front of giraffe 10.a young man stands on a white ledge and his food
Human Self-CIDEr: 0.967 **mBLEU-1: 0.557 mBLEU-2: 0.735** **mBLEU-3: 0.928 mBLEU-4: 1.000**	**SoftAtt-RS Self-CIDEr: 0.944** **mBLEU-1: 0.351 mBLEU-2: 0.654** **mBLEU-3: 0.843 mBLEU-4: 0.919**	**AdapAtt-RS Self-CIDEr: 0.936** **mBLEU-1: 0.395 mBLEU-2: 0.607** **mBLEU-3: 0.741 mBLEU-4: 0.948**

Figure 6. Captions generated by human, SoftAtt-RS and AdaptAtt-RS and the corresponding diversity scores, where a higher score indicates diverse captions. For the first image (top), human annotations obtain the lowest Self-CIDEr diversity score but highest mBLEU diversity scores. Looking at the captions, although human annotations use different words and phrases, they describe the same concept—*vase with flowers on a table*, whereas SoftAtt-RS describes not only *vase*, *flowers* and *table*, but also *leaves*, *window* and *house*, and AdapAtt-RS uses "clear", "water", "rainy light" and "green" to describe the image, which could be plausible descriptions. For the second image, both Self-CIDEr and mBLEU diversity metrics provide high scores and all captions almost use different words. Note that we use $-\log$ function in the Self-CIDEr diversity metric, which could be flatten and be less sensitive if the captions are relatively diverse (see figure 3).

1.a little girl is eating a chocolate doughnut
2.a small child eats a chocolate doughnut at a table
3.a little girl eating a chocolate frosted donut
4.a child is eating a choclate doughnout
5.a little girl enjoying a sweet confection and awaiting a sugar rush

Human LSA-based: 0.415 Self-CIDEr: 0.792

1.a girl that has a a both of her nose
2.a woman is talking on a cell phone
3.a woman with a small sleeping in her mouth
4.the blonde girl is black sweatshirt is holding a cell phone
5.a woman in a hotel dress sweat a peace
6.a woman is eating a cell phone and a shoe
7.a woman is eating a piece of cake with her nose
8.a girl is is brushing her teeth on a cell phone
9.a little girl is holding a cell phone
10.a girl on a cell phone on the bathrobe

CGAN-DRV LSA-based: 0.531 Self-CIDEr:0.859 acc: 0.230

1.a little girl eating a chocolate frosted donut
2.a close up of a person holding a banana
3.a small child eating a piece of food
4.a young baby is holding a piece of paper in it
5.a little girl holding a chocolate donut
6.a little girl eating a chocolate donut with sprinkles
7.a little girl eating a piece of bread
8.a girl is eating a banana on the street
9.a little girl eating a chocolate frosted donut
10.a small child is sitting on a table

GMMCVAE-DRV LSA-based: 0.499 Self-CIDEr: 0.732 acc: 1.255

1.a woman is holding a cell phone
2.a woman is talking on a cell phone
3.a woman is holding a cell phone
4.a woman is eating a cell phone
5.a woman is talking on a cell phone
6.a woman is holding a cell phone
7.a woman is holding a cell phone
8.a woman is talking on a cell phone
9.a woman is talking on her cell phone
10.a woman is talking on a cell phone

Att2in(C)-RS LSA-based: 0.189 Self-CIDEr: 0.358 acc: 0.070

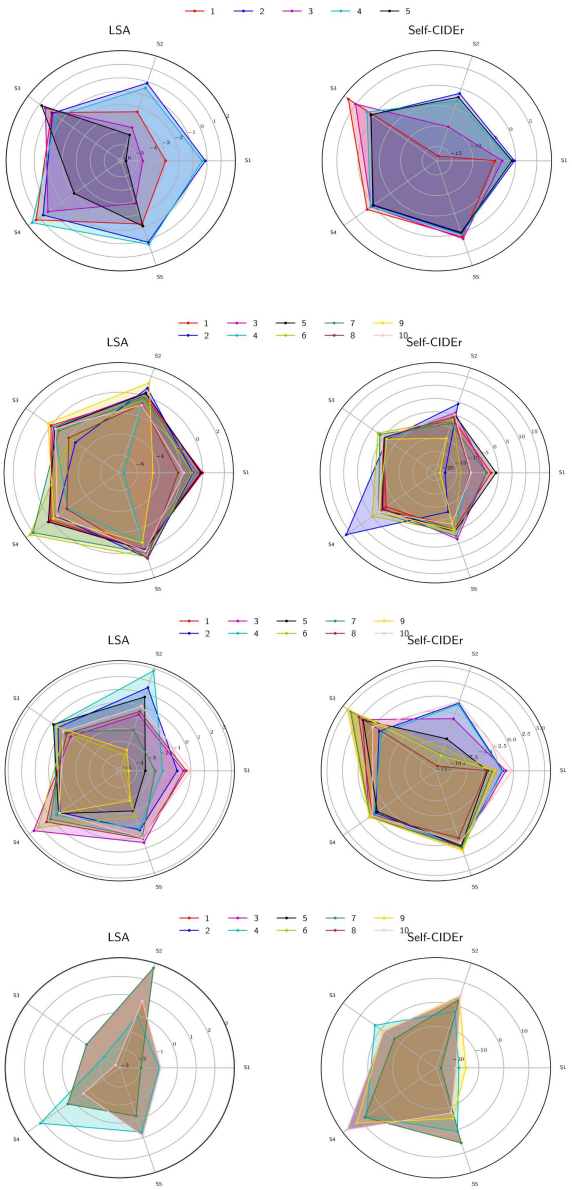

Figure 7. Visualizing captions in the latent semantic space. Human annotations focus on “little girl”, “eating” and “chocolate donut”, and looking at the radar chart of LSA, it roughly contains 2 semantics—S3 and S4. Moreover, captions 3 and 5 talk more about S3, and caption 5 does not talk about S4, if we compare captions 3 and 5, it is easy to find that both of them use “a little girl”, but caption 3 also uses “eating” and “doughnout”, therefore, S4 could denote “eating” somthing. Captions 6, 7 of CGAN, captions 3, 6 of GMMCVAE and caption 4 that use “eating” have a larger value of S4. Similarly, in the radar chart of Self-CIDEr, S3 could represent “eating” somthing and S4 could denote “talking”, thus the captions contain “eating” have relative large values of S3 and the captions that use “talking” could have larger values of S4.

1.a man sticking his head out of a doorway into a rainy city street
2.a man peeks out a window during a light rain
3.people are walking in the rain holding umbrellas
4.people walking outside in the rain under umbrellas and a man peeking his head out of a doorway
5.there are people walking down the street with umbrellas

Human LSA-based: 0.580 Self-CIDEr: 0.970

1.a group of people walking down in street on a rainy street
2.a group of people walking down the street with umbrellas
3.a group of people walking down a street holding umbrellas
4.many people walking at the city street their umbrellas
5.a group of people walking down a street holding umbrellas
6.a group of people walking in a rain soaked street
7.two men walking down a sidewalk under umbrellas
8.a group of people walking down a street with umbrellas umbrellas
9.a group of men walking down a rain soaked street
10.a group of people walking in the rain holding umbrellas

CGAN-DRV LSA-based: 0.431 Self-CIDEr:0.623 acc: 1.009

1.a man walking down a street with an umbrella
2.a man is walking down the street with an umbrella
3.consisting of a man in a room with a large umbrella
4.drum is in the middle of a street
5.people walking down the street with umbrellas
6.a man walking down a street with an umbrella
7.a man walking down a street with a black umbrella
8.a woman is eating a banana
9.solar street with a man riding a horse
10.a man is standing in the yard

GMMCVAE-DRV LSA-based: 0.485 Self-CIDEr: 0.723 acc: 0.434

1.a group of people walking in the rain with umbrellas
2.a group of people walking in the rain with umbrellas
3.a group of people walking in the rain with umbrellas
4.a group of people walking in the rain with umbrellas
5.a group of people walking in the rain with umbrellas
6.a group of people walking in the rain with umbrellas
7.a group of people walking in the rain with umbrellas
8.a group of people walking in the rain with umbrellas
9.a group of people walking in the rain with umbrellas
10.a group of people walking in the rain with umbrellas

Att2in(C)-RS LSA-based: 0.000 Self-CIDEr: 0.000 acc: 1.477

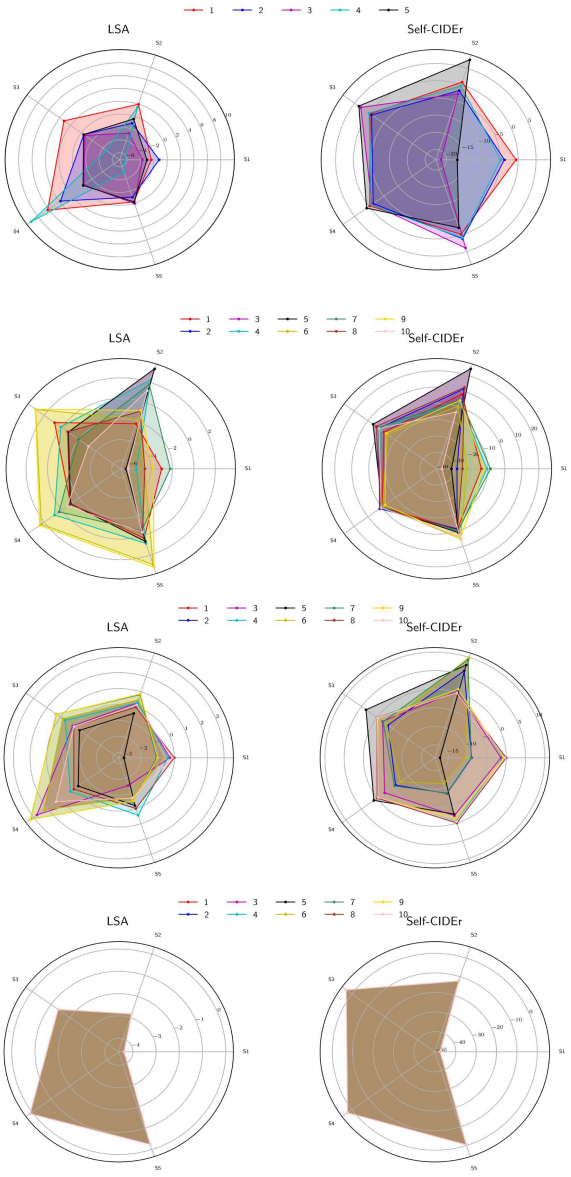

Figure 8. Visualizing captions in the latent semantic space.

1.a kid riding his skate board on the edge of a concrete wall
2.a man on a skateboard performs a trick on the ledge
3.a skateboarder doing a stunt off of a sidewalk planter
4.a man that is jumping a skateboard outside
5.a man on a skateboard performs tricks on the edge of a sidewalk planter

Human LSA-based: 0.381 Self-CIDEr: 0.824

1.a person on a skateboard does a trick on a skateboard
2.a person on the skateboard doing a trick near a ledge
3.a person riding a skateboard off of a pool
4.a skateboarder jumping a parking rail in the corner of a building
5.a person doing a skateboard on a city street
6.a man on a skateboard is jumping over big steps stairs
7.a man jumping a skateboard on a metal rail
8.a person doing a trick on a skateboard in front of
9.a person riding a skateboard in a city park
10.a man is doing a trick on a skateboard

CGAN-DRV LSA-based: 0.429 Self-CIDEr: 0.811 acc: 0.589

1.a skateboarder is doing a trick on a rail
2.a person on a skateboard on a ramp
3.two boy in the air with a skateboard
4.a person jumping in the air with a skateboard
5.a person on a skateboard does a trick
6.a young man is in the middle of a skateboard
7.a dog on a skateboard on a rail road
8.a person is doing a trick on a skateboard
9.a person on a skateboard jumping in the air
10.a a man on a skateboard is on a rail

GMMCVAE-DRV LSA-based: 0.417 Self-CIDEr: 0.793 acc: 0.553

1.a person is doing a trick on a skateboard
2.a person is doing a trick on a skateboard
3.a man is doing a trick on a skateboard
4.a man is doing a trick on a skateboard
5.a man is doing a trick on a skateboard
6.a person is doing a trick on a skateboard
7.a person is doing a trick on a skateboard
8.a person is doing a trick on a skateboard
9.a person is doing a trick on a skateboard
10.a man is doing a trick on a skateboard

Att2in(C)-RS LSA-based: 0.073 Self-CIDEr: 0.138 acc: 0.871

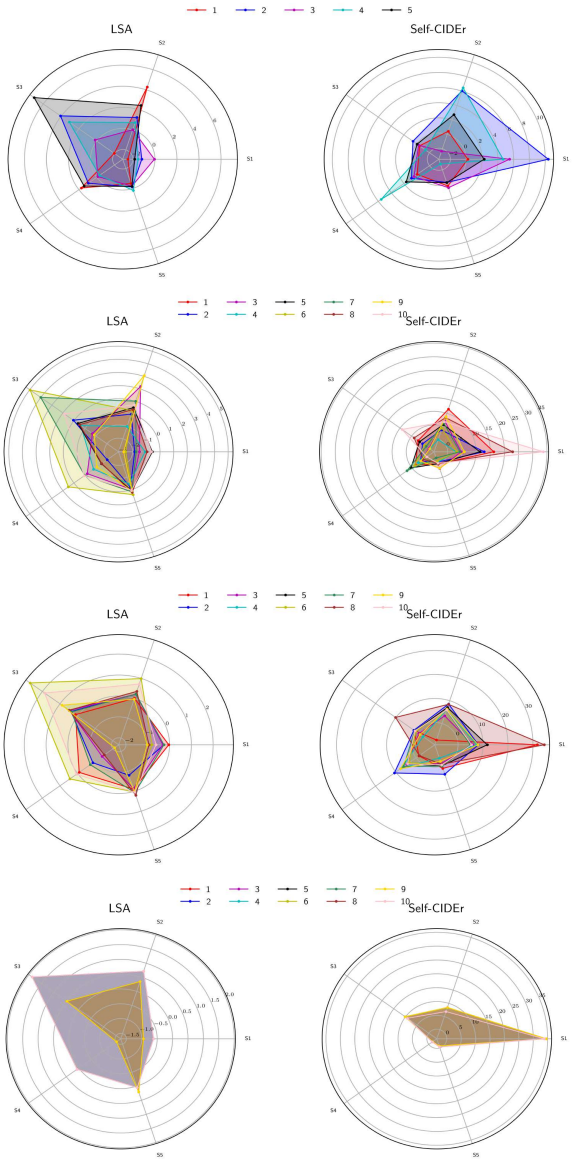

Figure 9. Visualizing captions in the latent semantic space.

1.woman standing in living room using video game controls
2.a woman standing in her living room playing wii
3.there is a woman that is ayi g with the wii in her room
4.a woman standing next to a couch holding a wii controller
5.young woman playing wii in a furnished living room

Human
Self-CIDEr: 0.854

1.there is a woman that is holding a wii video console
2.a girl a young girl are walking a video game
3.a woman standing in the living room playing a video game
4.a woman is playing wii in a living room
5.a beautiful about a pose while holding a video game controller
6.the person holding playing a game by herself
7.a woman is playing wii games while controller sets
8.a girls standing in a living room trying to use
9.girl woman in as the wii UNK and front of her girl wearing mouse
10.a holds with nintendo wii controller in hand

XE-only Self-CIDEr: 0.905 acc: 0.718

1.a girl in a room playing a video game
2.a woman standing in a living room with a wii remote
3.a woman plays video games in her living room
4.two women playing a video game in a living room
5.a woman holding a wii remote in a living room
6.a woman is playing the wii in the room room
7.a woman playing a game remote in no room
8.a girl playing a wii for her wii
9.a woman is standing a video game in a living room
10.a woman is standing his wooden in a living room

lambda=5 Self-CIDEr: 0.808 acc: 0.944

1.a woman standing in a living room holding a nintendo wii controller
2.a woman standing in a living room playing with a controller
3.a woman standing in a living room playing a video game
4.a beautiful attractive girl playing a video game
5.a woman standing in a living room holding a video game controller
6.a woman standing in a living room playing a video game
7.a girl playing a video game in a living room
8.a woman standing in a living room playing a video game
9.a girl standing on a couch playing a video game
10.a woman standing in a living room playing a video game

lambda=10 Self-CIDEr: 0.513 acc: 1.206

1.a woman standing in a living room playing a video game
2.a woman standing in a living room playing a video game
3.a woman standing in a living room playing a video game
4.a woman standing in a living room playing a video game
5.a woman standing in a living room holding a video game
6.a woman standing in a living room playing a video game
7.a woman standing in a living room playing a wii game
8.a woman standing in a living room playing a video game
9.a woman standing in a living room playing a video game
10.a woman standing in a couch playing a video game

lambda=20 Self-CIDEr: 0.306 acc: 1.484

1.a woman playing a video game in a living room
2.a woman standing in a living room playing a video game
3.a woman standing in a living room playing a video game
4.a woman standing in a living room playing a video game
5.a woman playing a video game in a living room
6.a woman playing a video game in a living room
7.a woman playing a video game in a living room
8.a woman standing in a living room playing a video game
9.a woman holding a video game in a room
10.a woman standing in a living room playing a video game

CIDEr-only Self-CIDEr: 0.273 acc: 1.296

1.a woman standing playing a video game in a living room
2.a woman standing in a living room with a video game
3.a woman standing in a living room holding a video game
4.a woman standing playing a video game in a living room
5.a woman standing playing a video game in a living room
6.a woman standing holding a video game in a living room
7.a woman standing in a living room playing a video game
8.a woman standing enjoying a living room with a couch
9.a woman standing in a living room with a video game controller
10.a woman standing in a living room with a video game

beta=10 Self-CIDEr: 0.472 acc: 1.278

Figure 10. Captions generated by Att2in model traind with different conbinations of loss functions. **lambda** denotes the weight of CIDEr reward and **beta** denotes the weight of retrieval reward (see section 5.3 in our paper). Using large **lambda** is able to increase the accuracy but reduce diversity.

1.a pot full of beef and broccoli stew
2.a broccoli and beef dish with baby corn
3.a pot of food contains meats and vegetables
4.soup with broccoli and meat cooking on a stove
5.broccoli and meat in a large pot that is ready for serving

Human
Self-CIDEr: 0.949

1.a stir fry dish fry being cooked and broccoli
2.a mixture of stir fry on a UNK
3.a pitcher filled with lots broccoli and in a stove
4.a black plate filled fresh cooked with veggies
5.broccoli over a pan of with broccoli bowl
6.a pan full of vegetables in sauce pot
7.a plate of meat broccoli broccoli a row
8.food are being to be group of vegetables in a bowl
9.this dish of broccoli has broccoli and top
10.a dish of meat sits meat on the table

XE-only Self-CIDEr: 0.918 acc: 0.290

1.a plate of food with carrots and broccoli on
2.a close up of several different kinds of stew and broccoli
3.a skillet of steamed broccoli and white vegetable
4.a stir fry of broccoli onions carrots and mashed chicken
5.a bowl of broccoli containing cauliflower and a large bin on the
6.a bowl of noodles broccoli and broccoli being stirred
7.a plate topped with with potatoes and and broccoli
8.a pan filled with meat and broccoli and broccoli on a counter
9.a stir fry dish topped with meat and broccoli
10.broccoli on of and beans and tofu on a counter

lambda=5 Self-CIDEr: 0.881 acc: 0.394

1.a plate of broccoli and noodles that are being cooked
2.a plate of broccoli and broccoli are on a table
3.a plate of creamy broccoli and asparagus on a table
4.a plate of broccoli and vegetables sitting on a table
5.a bowl of broccoli and broccoli on a table
6.a meal of broccoli carrots and seasoning on a
7.a pan of food and broccoli in a counter
8.a plate of of chinese food and vegetables on a stove
9.a square dinner with broccoli and tofu on the counter
10.a plate with a mixture of broccoli and and vegetables

lambda=10 Self-CIDEr: 0.812 acc: 0.473

1.a bowl of food and broccoli on a stove
2.a plate of food chicken broccoli broccoli and broccoli
3.a plate of pasta broccoli and broccoli on a white
4.a bowl of broccoli and broccoli on a stove
5.a close up of a bowl of broccoli broccoli and chicken
6.a plate of pasta broccoli and noodles on a counter
7.a plate of food and broccoli on a table
8.a plate of food and broccoli on a table
9.a plate of broccoli and broccoli on a stove
10.a pan of broccoli and broccoli on a table

lambda=20 Self-CIDEr: 0.608 acc: 0.568

1.a plate of food and broccoli on a stove
2.a plate of food and broccoli on a stove
3.a plate of food and broccoli on a stove
4.a bowl of broccoli and vegetables on a stove
5.a plate of broccoli and vegetables on a stove
6.a plate of food and broccoli on a stove
7.a plate of food and broccoli on a stove
8.a plate of food and broccoli on a stove
9.a bowl of broccoli and vegetables on a stove
10.a bowl of broccoli and broccoli on a stove

CIDEr-only Self-CIDEr: 0.318 acc: 0.820

1.a bowl of broccoli and broccoli cooking in a pan
2.a pan of broccoli and broccoli in a pan
3.a pan of broccoli and noodles mushrooms on a pan
4.a pan of broccoli broccoli and vegetables in a pan
5.a pan of broccoli and vegetables in a pan
6.a pan of broccoli and vegetables in a pan
7.a pan of broccoli cauliflower and broccoli on a pan
8.a pan of food noodles broccoli and meat on a pan
9.a pan of broccoli and broccoli on a pan
10.a pan of broccoli stew and mushrooms broccoli in a pan

beta=10 Self-CIDEr: 0.567 acc: 0.444

Figure 11. Captions generated by Att2in model traind with different conbinations of loss functions. **lambda** denotes the weight of CIDEr reward and **beta** denotes the weight of retrieval reward (see section 5.3 in our paper).