Partial Off-policy Learning: Balance Accuracy and Diversity for Human-Oriented Image Captioning - Supplementary Material

Jiahe Shi Yali Li* Shengjin Wang
Department of Eletronic Engineering, Tsinghua University, Beijing 100084, China
shijh18@mails.tsinghua.edu.cn {liyali13, wgsgj}@mail.tsinghua.edu.cn

1. Accuracy-Diversity Balance of Human Performance

In our paper, we evaluate human performance based on the annotations within the MSCOCO dataset. To be specific, we follow the implementation in [16, 17] where the leave-one-out CIDEr score is used as an evaluation for accuracy. Mathematically, the score is calculated as:

$$\text{CIDEr}(\mathcal{G}(I)) = \frac{1}{|\mathcal{G}(I)|}\sum_{s\in\mathcal{G}(I)}\text{CIDEr}^*(s), \quad \text{where}$$

$$\text{CIDEr}^*(s) = \frac{1}{|\complement\mathcal{G}(I)\{s\}|}\sum_{j\in\complement\mathcal{G}(I)\{s\}}\cos\langle\mathbf{g}_n(s),\mathbf{g}_n(j)\rangle$$

We notice that human acquires CIDEr score far lower than the state-of-the-art image captioning methods. This issue may be caused by some inaccurate annotations within the dataset. Recently, Wang *et al.* [18] reports a similar issue that some of the annotations are of "low qualities", *i.e.* acquire CIDEr score significantly lower than other annotations of the image. To provide a comprehensive and fair evaluation of human performance, we exclude the offbeat annotations from evaluating human performance. Specifically, we calculated $\text{CIDEr}^*(s)$ for each $s \in \mathcal{G}(I)$. The annotation s with relatively lower $\text{CIDEr}^*(s)$ among the 5 ground truths $\mathcal{G}(I)$ is regarded as the semantically different one. By excluding the most or the most two different annotations of the image, we derive a curve indicating human performance on accuracy-diversity balance in Fig. 1.

In Fig. 1, we gradually exclude the most different annotations according to the leave-one-out CIDEr score. With offbeat sentences excluded from the evaluation, the remaining sentences become semantically similar to each other, resulting in high accuracy scores and relatively low diversity performance. Note that our method with ϵ set as 0.1 still locates the closest to the curve of human performance, which means we achieve better human-oriented performance in terms of accuracy and diversity.

*corresponding author

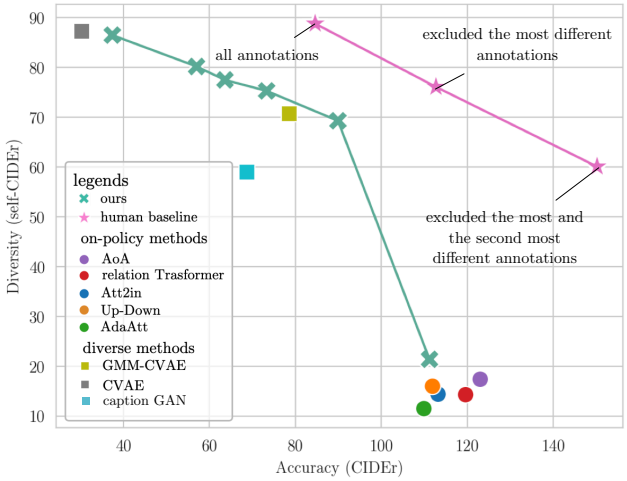


Figure 1. Accuracy-Diversity balance of human performance by excluding certain offbeat annotations. Model trained by the proposed partial off-policy strategy with ϵ set as 0.1 is still the closest to the curve of human performance.

	$\|\rho_{all}\|$	$\|\rho_{ex\text{-}1}\|$	$\|\rho_{ex\text{-}2}\|$
Att2in[10]	0.243	0.189	0.132
Up-Down[2]	0.253	0.200	0.139
AdaAtt[9]	0.236	0.181	0.123
ReTrans[6]	0.295	0.238	0.170
AoA[8]	0.289	0.232	0.171
CVAE[15]	0.200	0.168	0.138
GMM-CVAE[15]	0.280	0.181	0.129
CapGAN[11]	0.193	0.157	0.102
Ours ($\epsilon = 0.1$)	**0.337**	**0.304**	**0.255**

Table 1. Pearson's correlation coefficient between predicted captions and subsets of human annotations, where ex-1 denotes excluding the most different annotations from evaluation and ex-2 denotes excluding the most two different annotations.

Accordingly, we also report Pearson's correlation coefficient between model-generated captions and the subset of human annotations in Table 1. Our method achieves correlation coefficients as 0.304 and 0.255 respectively excluding

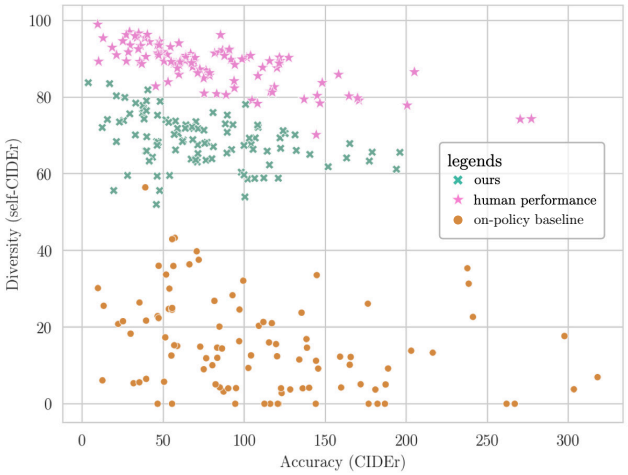


Figure 2. The distribution of the performance acquired by different methods in the Diversity-Accuracy space. We sample 100 images randomly from Karpathy's validation set for visualization. Compared with the on-policy trained baseline, our approach not only locates closer to human performance but also manages to mimic the human distribution.

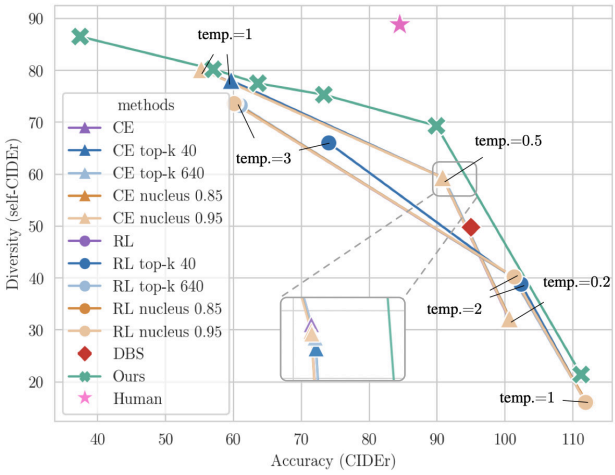


Figure 3. Performance compared with sampling methods. All the performances are acquired using Top-down model [2].

the most or the most two different annotations, which is still the best performance compared with other methods.

The curve of human performance demonstrated in Fig. 1 also acts as the performance boundary for which works on the accuracy-diversity balance issue are supposed to approach. Currently, diversity performance derived by the proposed partial off-policy strategy declines significantly when accuracy increases. However, since our work is agnostic to model structure, we would expect to narrow the gap by deploying the proposed learning scheme on models with better accuracy in the future.

2. What does Pearson's Correlation Coefficient Stand for?

We use Pearson's correlation coefficient as the quantitative evaluation in our paper. It actually serves as a complement to the diversity-accuracy diagram. In the diagram, we locate closest to human performance, indicating that we approach human performance in general. However, we are also concerned about the quality of the modeled posterior in detail, as illustrated in Fig. 2.

By acquiring a higher Pearson's correlation coefficient, the performance of our method varies over different image inputs in a similar manner compared with humans. In other words, models trained by the proposed scheme behave well where humans give better annotations and behave relatively poorly where humans fail to provide decent captions too. **This indicates that we fully utilize the supervision annotations and succeed in modeling a human-like posterior**.

Furthermore, we would like to emphasize that the correlation coefficient cannot represent the quality of the posterior by itself. It is supposed to be combined with the diversity-accuracy diagram for a comprehensive evaluation, where the diagram illustrates that our method achieves human-like performance from a holistic perspective, and Pearson's correlation coefficient illustrates such a conclusion considering the detail characteristic of the posterior.

3. Additional Quantitative Results

More Comparison Results with Sampling Methods. Our method is a training-side improvement which differs from sampling methods like temperature scaling, top-k sampling [4], nucleus sampling [7], and DBS [14]. In our paper, we provide a comparison with DBS [14]. We would like to provide more comparison results with the above-mentioned sampling strategies on the balance effect of accuracy and diversity. The result is demonstrated in Fig. 3, where different sampling methods are performed on both CE-trained model and on-policy RL-trained model.

Our method forms an upper envelope over the sampling approaches in Fig. 3, which suggesting better performance. This is because that the sampling methods achieve balance effects based on manipulation over a learned posterior, while our method manages to derive a posterior of better quality with the training-side improvements.

Comparison with Other On-policy Exploration Methods. Compared with the mainstream on-policy approaches in image captioning, our method can be viewed as an exploration-enhancing strategy. There is also some literature focusing on the exploration issue under the on-policy framework, *i.e.* distributional entropy regularization [5] and count-based state visitation [12]. However, as we mentioned in our paper, such traditional exploration strategies may not perform well in image captioning. To illustrate

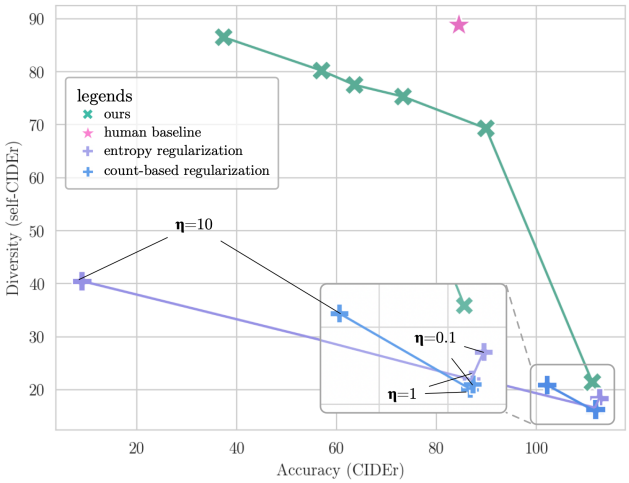


Figure 4. Performance compared with entropy regularization and count-based visitation approaches. We use a regulation term η to control the exploration intensity, where higher η indicates more intensive exploration. For the count-based method, we follow [12] to use 32-bit sim-hash for the representation of the states.

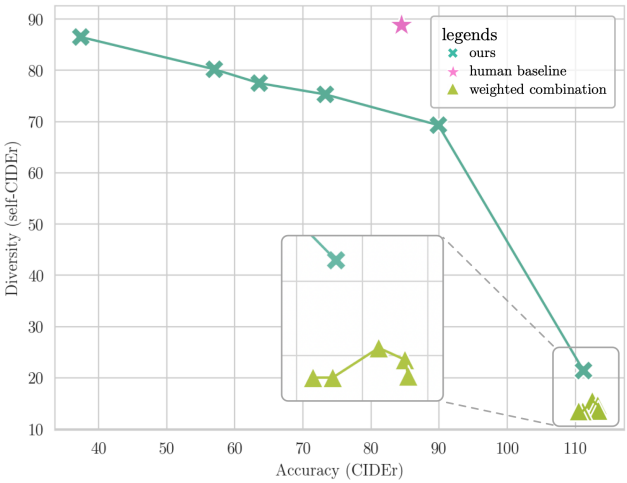


Figure 5. Performance compared with weighted combination solution in multi-objective reinforcement learning. Our method achieves better performance due to explicitly modeling the correspondence between rewards and policies.

such an issue, we conduct experiments and observe performance degradation on both diversity and accuracy. The experimental result can be seen in Fig 4.

The ineffectiveness of traditional exploration methods may due to the enormous searching space in the task of image captioning. The action space of image captioning in each step consists of 9,487 different actions (words). Typically the maximum length of each trajectory is set as 16. Therefore there are up to $9487^{16} \approx 4.3 \times 10^{63}$ different states within the searching space. Simply encouraging exploration indiscriminately in such searching space may be inefficient and hinder the model from acquiring better performance. On the contrary, our method restricts the searching space by introducing the sampling model, which circumvents such an issue.

Does Weighted Combination of Reward Components Works? As we mentioned in our paper, the proposed *partial off-policy* learning scheme can be interpreted under the scheme of multi-objective reinforcement learning. An ordinary solution to such problems is to use a weighted combination of the multiple rewards as the optimization target. We conduct experiments to provide a comparison between our method and the above-mentioned solution. The results can be found in Fig 5.

Our method acquires better performance than the weighted combination solutions. This is because in our work we not only optimize the additional reward *max-CIDEr* but also arrange a specific way (*i.e.* the off-policy strategy) to generate appropriate training trajectories. Thus, the goal is optimized more effectively.

Other combination of diversity and accuracy evaluation. In our paper, we evaluate the proposed method using CIDEr [13] as accuracy evaluation and self-CIDEr [16] as diversity evaluation. We would like to present more evaluation results with different combinations of diversity-accuracy metrics for comprehensive illustration. For accuracy scores, we report METEOR [3] in addition to CIDEr, for its comprehensive evaluation on both recall and precision. We also report SPICE [1] for its insensitivity on n-gram overlap. For diversity evaluations, we report self-CIDEr and mBLEU-4. Since lower mBLEU-4 represents more diversity, we reverse it as 100 − mBLEU-4 so that it is consistent with self-CIDEr. Moreover, since mBLEU-4 varies dramatically, we plot the figure using logarithmic coordinates for more details. The results are illustrated in Fig. 6, where our method locates close to human performance according to all metric selections.

We also compute the Pearson's correlation coefficient using different combinations of metrics. The results are shown in Table 2. We achieve the best correlation with human scores according to most selections of metrics. For example, our method obtains correlation coefficients as 0.324, 0.260, and 0.333 respectively using METEOR, BLEU-4, and SPICE for accuracy evaluation and self-CIDEr as diversity metric, which exceeds the performance of other methods by a significant margin. We also notice that some methods outperform ours when mBLEU-4 is chosen to represent diversity. This may due to the significant variation of mBLEU-4 as reported by [16]. Since we normalize the diversity and accuracy scores respectively before calculating correlation coefficients, the influence of diversity may be greatly reduced due to the large variance of mBLEU-4, leading the coefficients to incline towards accuracy more. To provide fair evaluation, we calculate the correlation co-

	$\|\rho_{\mathrm{C+sC}}\|$	$\|\rho_{\mathrm{C+mB}}\|$	$\|\rho_{\mathrm{M+sC}}\|$	$\|\rho_{\mathrm{M+mB}}\|$	$\|\rho_{\mathrm{B+sC}}\|$	$\|\rho_{\mathrm{B+mB}}\|$	$\|\rho_{\mathrm{S+sC}}\|$	$\|\rho_{\mathrm{S+mB}}\|$
Att2in[10]	0.243	0.417	0.180	0.268	0.142	0.201	0.174	0.316
Up-Down[2]	0.253	0.429	0.189	0.284	0.150	0.218	0.177	0.317
AdaAtt[9]	0.236	0.396	0.168	0.245	0.129	0.179	0.162	0.300
ReTrans[6]	0.295	0.472	0.220	0.311	0.182	0.250	0.214	**0.362**
AoA[8]	0.289	**0.478**	0.221	0.317	0.177	0.246	0.214	0.357
CVAE[15]	0.200	0.185	0.174	0.158	0.141	0.121	0.179	0.163
GMM-CVAE[15]	0.280	0.255	0.178	0.190	0.124	0.131	0.181	0.200
CapGAN[11]	0.193	0.282	0.168	0.221	0.109	0.139	0.139	0.189
Ours ($\epsilon = 0.1$)	**0.337**	0.379	**0.324**	**0.336**	**0.260**	**0.264**	**0.333**	0.344

Table 2. Comprehensive correlation coefficient between predicted captions and human annotations using different combinations of evaluation metrics, where C denotes CIDEr, M denotes METEOR, B denotes BLEU-4, S denotes SPICE, sC denotes self-CIDEr and mB denotes mBLEU-4.

	$\|\rho_{\mathrm{C}}\|$	$\|\rho_{\mathrm{M}}\|$	$\|\rho_{\mathrm{B}}\|$	$\|\rho_{\mathrm{S}}\|$
Att2in[10]	0.388	0.250	0.187	0.300
Up-Down[2]	0.399	0.267	0.203	0.299
AdaAtt[9]	0.372	0.233	0.168	0.288
ReTrans[6]	0.444	0.296	0.233	0.343
AoA[8]	**0.445**	0.300	0.228	0.339
CVAE[15]	0.218	0.158	0.127	0.167
GMM-CVAE[15]	0.356	0.230	0.172	0.278
CapGAN[11]	0.358	0.249	0.150	0.217
Ours ($\epsilon = 0.1$)	**0.445**	**0.332**	**0.264**	**0.371**

Table 3. Comprehensive correlation coefficient between predicted captions and human annotations using different combinations of evaluation metrics. The diversity performance is represented by $\exp(100-$ mBLEU-4$)$, which is consistent with Fig. 7.

efficient using $\exp(100 - \text{mBLEU-4})$ in Table 3, as the logarithmic coordinates tends to demonstrate more details in Fig. 7. Our method achieves the best performance compared with other works under such implementation.

4. Additional Qualitative Results

We also present more samples for qualitative evaluation. The results are shown in Fig. 7. Descriptive semantics which the on-policy baseline omits are **highlighted**. Moreover, we provides some samples in the last row of Fig. 7 where incorrect descriptions are predicted, which results in decline of accuracy performance. The incorrect concepts include false attribute (*e.g.* "white" of the left image), false relation (*e.g.* "reach" of the middle image) and false object (*e.g.* "carrot" of the right image). We expect future works to deal with such cases.

References

[1] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European Conference on Computer Vision*, pages 382–398. Springer, 2016. 3

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. 1, 2, 4

[3] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380, 2014. 3

[4] Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, 2018. 2

[5] Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. Reinforcement learning with deep energy-based policies. In *International Conference on Machine Learning*, pages 1352–1361. PMLR, 2017. 2

[6] Simao Herdade, Armin Kappeler, Kofi Boakye, and Joao Soares. Image captioning: Transforming objects into words. In *Advances in Neural Information Processing Systems*, pages 11137–11147, 2019. 1, 4

[7] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2019. 2

[8] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4634–4643, 2019. 1, 4

[9] Jiasen Lu, Caiming Xiong, Devi Parikh, and Richard Socher. Knowing when to look: Adaptive attention via a visual sentinel for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 375–383, 2017. 1, 4

[10] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for

image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7008–7024, 2017. 1, 4

[11] Rakshith Shetty, Marcus Rohrbach, Lisa Anne Hendricks, Mario Fritz, and Bernt Schiele. Speaking the same language: Matching machine to human captions by adversarial training. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4135–4144, 2017. 1, 4

[12] Haoran Tang, Rein Houthooft34, Davis Foote, Adam Stooke, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. # exploration: A study of count-based exploration for deep reinforcement learning. 2, 3

[13] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015. 3

[14] Ashwin K Vijayakumar, Michael Cogswell, Ramprasaath R Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. Diverse beam search for improved description of complex scenes. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 2

[15] Liwei Wang, Alexander Schwing, and Svetlana Lazebnik. Diverse and accurate image description using a variational auto-encoder with an additive gaussian encoding space. In *Advances in Neural Information Processing Systems*, pages 5756–5766, 2017. 1, 4

[16] Qingzhong Wang and Antoni B Chan. Describing like humans: on diversity in image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4195–4203, 2019. 1, 3

[17] Qingzhong Wang, Jia Wan, and Antoni B Chan. On diversity in image captioning: Metrics and methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1

[18] Ziwei Wang, Zi Huang, and Yadan Luo. Human consensus-oriented image captioning. In *Proc. International Joint Conference on Artificial Intelligence, IJCAI*, pages 659–665, 2020. 1

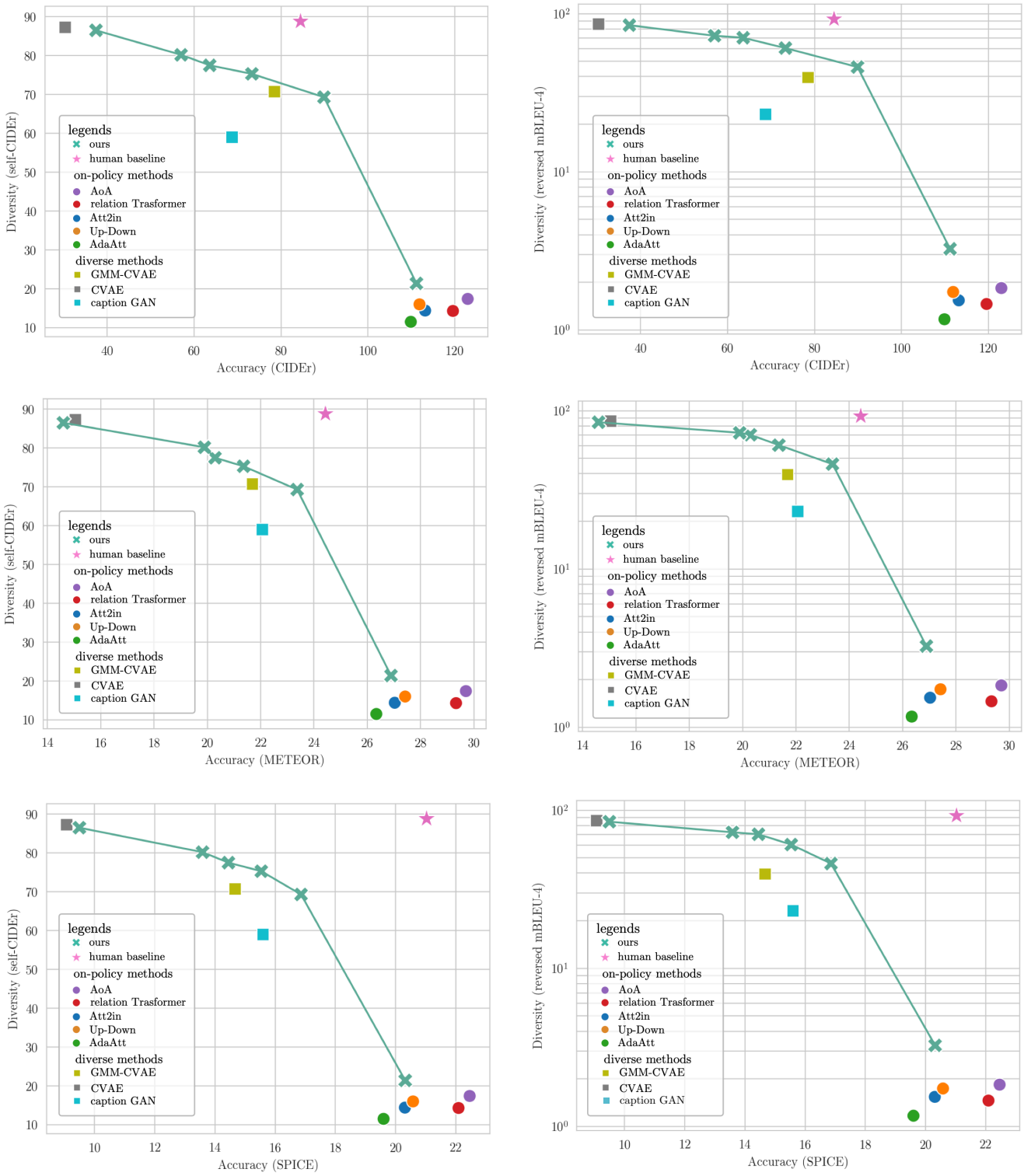
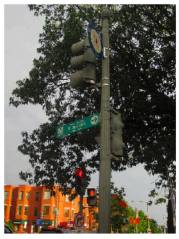


Figure 6. Accuracy-Diversity balance according to multiple combinations of metrics. Our method shows human-oriented performance over different metric selections.



On-policy Baseline:
- A street sign on the side of a traffic light.
- A street sign on the side of a traffic light.
- A street sign on the side of a traffic light.
- A street sign on the side of a traffic light.
- A street sign on the side of a traffic light.

Ours:
- A traffic light sitting next to a **street**.
- A traffic light on a pole **in front of a building.**
- A street sign posted on a traffic light.
- A street sign is in front of a tree **in a city**.
- A street sign sitting **in front of a building** next to trees.

Human:
- A red building standing against a grey sky.
- A multi-purpose traffic signal street sign structure positioned along the street.
- An intersection with a stoplight has an orange building in the background.
- A traffic light and street sign in a city.
- A traffic light with a street sign in front of a large orange building.

On-policy Baseline:
- Two men playing with a frisbee in a field.
- Two men playing with a frisbee in a field.
- Two men playing with a frisbee in a field.
- Two men playing with a frisbee in a field.
- Two men playing with a frisbee in a field.

Ours:
- A couple of men playing a game of frisbee.
- A group of men playing frisbee in a field.
- A couple of men playing frisbee in the **grass**.
- **Shirtless** men playing a game of frisbee in an open field.
- Two people playing a game of frisbee **in a park**.

Human:
- Two guys are playing frisbee in the park.
- Two young men playing a shirtless game of frisbee.
- Two men in a grassy field playing with a frisbee.
- Two men are playing with a frisbee together.
- Two shirtless men playing frisbee in a field.

On-policy Baseline:
- A herd of zebras and a zebra standing in a field.
- A herd of zebras and a zebra standing in a field.
- A herd of zebras and a zebra standing in a field.
- A herd of zebras and a zebra standing in a field.
- A herd of zebras and a zebra standing in a field.

Ours:
- A herd of zebras are grazing in a field.
- A herd of zebras grazing on grass in a **park**.
- A herd of zebras and **giraffes** grazing in a **grassy** area.
- Some zebra and some **giraffes** in a field.
- There are a number of zebras and some **giraffes** out in the wild.

Human:
- These four zebra are walking in a field.
- Four zebras walking in a grassy area.
- A herd of zebra walking along a lush green field.
- A group of zebra eating grass in a field near a couple of giraffe.
- Four zebras and three giraffes wander in a preserve.

On-policy Baseline:
- A woman standing in front of a birthday cake with candles.
- A woman standing in front of a birthday cake with candles.
- A woman standing in front of a birthday cake with candles.
- A woman standing in front of a birthday cake with candles.
- A woman standing in front of a birthday cake with candles.

Ours:
- A woman standing in front of a birthday cake with candles.
- A woman standing in front of a birthday cake with lit candles.
- A woman about to **blow out** candles on a birthday cake.
- A young female **holds** a cake with a candle on the top of it.
- A woman is standing in front of a birthday cake.

Human:
- A woman handing another woman a birthday cake filled with candles.
- A woman holding a blue birthday cake with stars and candles on it and another woman in front of the cake.
- A women recieves a cake that is blue.
- A fat girl blowing out candles on a cake.
- One lady is holding a birthday cake while another blows out the candles.

On-policy Baseline:
- A group of people playing soccer on a field.
- A group of people playing soccer on a field.
- A group of people playing soccer on a field.
- A group of men playing soccer on a field.
- A group of men playing soccer on a field.

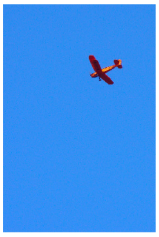
Ours:
- A group of people playing soccer on a field.
- A group of people playing soccer on a field.
- A group of men playing a game of soccer.
- Several men playing soccer and **a crowd watching**.
- **People watching** athletes play a game of soccer.

Human:
- A group of men playing a game of soccer on a field.
- A crowd is watching a soccer game being played.
- Soccer game in process with fans behind to support the game.
- People are outside in a field playing with a soccer ball.
- Men playing soccer during a game at a stadium.

On-policy Baseline:
- A building with a clock tower on top of it.
- A building with a clock tower on top of it.
- A building with a clock tower on top of it.
- A building with a clock tower on top of it.
- A building with a clock tower on top of it.

Ours:
- A clock tower on top of a building.
- A large **stone** building with a clock on the front.
- A large building with a clock on the top.
- A brick building with a clock on the front of it.
- A clock tower on the side of a building.

Human:
- A medieval style tower and clock against blue sky.
- A stone clock tower sits beneath a cloudy blue sky.
- A clock sits atop a tower claiming the time to be 2:56.
- A historic clock tower turret still keeps the time.
- A stone clock tower with a wind vane on its steeple.



On-policy Baseline:
- A red airplane is flying in the blue sky.
- A red airplane is flying in the blue sky.
- A red airplane is flying in the blue sky.
- A red airplane is flying in the blue sky.
- A red plane is flying in the blue sky.

Ours:
- A red **and white** airplane is flying in the sky.
- An airplane that is flying in the sky.
- A red plane flying in the air.
- A plane flying in the sky **on a sunny day**.
- A plane is flying through **a clear sky**.

Human:
- A red plane flying in a cloudless sky.
- A small airplane flying through a blue sky.
- A small airplane flies in the sky during the daytime.
- A small plane flying in the air with a sky background.
- A red plane flying high in a clear blue sky.

On-policy Baseline:
- A man **holding a tennis ball** on a tennis court.
- A man **holding a tennis ball** on a tennis court.
- A man **holding a tennis ball** on a tennis court.
- A man **holding a tennis ball** on a tennis court.
- A man **holding a tennis ball** on a tennis court.

Ours:
- A man standing on a tennis court holding a racquet.
- A man **is reaching to hit** a tennis ball on a court.
- A young boy **playing** tennis on a tennis court.
- A man is **swinging** a tennis racket on a court.
- A boy is in the middle of the tennis court.

Human:
- A young boy getting ready to swing at a tennis ball.
- A boy about to hit a tennis ball.
- Young boy playing tennis gets ready to hit the ball.
- A young boy on a tennis court with a racket.
- Boy poised about to forehand smash a tennis ball.

On-policy Baseline:
- A plate of food with broccoli and vegetables.
- A plate of food with broccoli and vegetables.
- A plate of food with broccoli and vegetables.
- A plate of food with broccoli and vegetables.
- A plate of food with vegetables **on a table**.

Ours:
- A close up of a plate of food with broccoli.
- A close up of a dish of broccoli and **carrots**.
- A large plate of vegetables and **tomatoes** on a <*unk*>.
- A **meat and grain pizza** with cut vegetables.
- Broccoli with the **trimmings ready to be cut in**.

Human:
- A vegetable and cheese pizza with broccoli and tomatoes.
- A meal made from vegetables, tomatoes, and spices.
- A pizza is topped with spinach, tomatoes, and broccoli.
- A meal containing spinach, tomatoes, cheese, and broccoli.
- Broccoli, tomatoes, and spinach on a bed of melted cheese.

Figure 7. More qualitative results.