SC-Captioner: Improving Image Captioning with Self-Correction by Reinforcement Learning

Supplementary Material

8. Prompt Templates

We follow the official instruction of each LVLM and adopt simple prompt for image captioning and self-correction. For relation evaluation, we prompt open-source language models to answer the given questions based on the candidate captions. The Prompts used are illustrated as follows:

Prompt For LLaVA-1.5

Captioning:

- A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions. USER: Caption this image as accurately as possible, without speculation. Describe what you see. ASSISTANT:

Self-Correction:

- <History> USER: The previous response is not very good. Please review the objects, attributes and relations in the caption. Remove that not appear in the image and add missing ones in the previous caption. Directly output the final caption: ASSISTANT:

Prompt For Qwen2-VL

Captioning:

- system\nYou are a helpful assistant.\nuser\nCaption this image as accurately as possible, without speculation. Describe what you see.\nassistant\n

Self-Correction:

- <History>user\nThe previous response is not very good. Please review the objects, attributes and relations in the caption. Remove that not appear in the image and add missing ones in the previous caption. Directly output the final caption: \nassistant\n

Figure 5. Prompts for image captioning and self-correction.

Prompt For Relation QA

- I will give you a passage of caption. Please answer the following 5 questions with \"Yes\", \"No\", or \"n/a\" based on the given caption. Output like this: \"1: Yes, 2: No, 3: Yes, 4: n/a, 5: Yes\". Don't output extra text.
  Caption: "<Caption>"
  Questions:1. <Question1>2. <Question2>3. <Question3>4. <Question4>5. <Question5>

Figure 6. Prompts for Relation evaluation via QA.

9. Human Consistency of Proposed Metric

We conducted an extra experiment to investigate how well our proposed metric aligns with human judgement. We randomly select 100 images in DOCCI500 and ask 4 human annotators to sort the captions provided by 4 different models, while considering both precision and recall. Then we calculate the Kendall's τ of BLEU-4, METEOR, CAPTURE and our metric (weighting 5,5,2 for objects, attributes and relations). The results in Tab. 4 show that our metric has better alignment with human judgement.

	BLEU-4	METEOR	CAPTURE	Our Metric
Kendall's τ	30.73	32.26	37.66	**45.99**

Table 4. Correlation of metrics and human judgments. Our metric gets higher score than CAPTURE and traditional metrics.

10. Statistics of Captions

We have made some analyses on different datasets including RefinedCaps, DOCCI, DCI and Localized Narratives in Tab. 5. As shown in the table, captions in our proposed dataset are relatively long and have more densely packed descriptions about objects, attributes and relations.

Dataset	Words	Objects	Attributes	Relations
RefinedCaps	120.53	**16.82**	**16.14**	**11.90**
DOCCI	121.91	13.33	14.29	10.50
DCI	**133.23**	15.90	14.14	10.90
Localized Narratives*	40.47	6.89	1.52	4.45
COCO-LN500	77.46	11.26	2.89	7.68

Table 5. Statistics across different datasets. * denotes that only a subset on COCO is selected. RefinedCaps has the highest element density.

11. Additional Experiments

11.1. Results of Using Public Dataset for Training

We also use the training set of DOCCI [33] which consists of 9.7K image-caption pairs as the training set for supervised fine-tuning and self-correction training of Qwen2-VL. Metrics for both the initial and self-corrected captions are shown on both DOCCI500 and COCO-LN500 datasets in Tab. 6. It can be seen that our proposed method outperforms SFT and DPO by a considerable margin, especially in crucial F1 and QA metrics, demonstrating the universality of our method.

Scenario	Post-training	BLEU-4	METEOR	CAPTURE	Objects			Attributes			Relations
					Precision	Recall	F1	Precision	Recall	F1	QA
Same-Domain	SFT	40.29	25.44	62.53	78.01	65.30	70.31	67.13	49.33	56.08	25.43
	SFT*	41.78	26.13	62.72	77.70	65.87	70.52	66.99	49.94	56.47	25.96
	SFT+DPO	42.79	25.90	63.04	76.95	66.65	70.71	66.44	50.17	56.40	27.25
	SFT+DPO*	**43.20**	**26.80**	63.28	75.53	**67.90**	71.02	64.52	**51.34**	56.46	27.82
	SFT+Ours	41.10	26.11	63.47	**79.09**	66.30	71.45	**70.09**	50.04	57.63	26.68
	SFT+Ours*	41.95	26.38	**63.83**	78.85	67.59	**72.05**	69.77	50.64	**58.00**	**28.58**
Cross-Domain	SFT	**34.93**	25.97	47.63	78.03	70.42	73.43	67.91	53.51	56.31	30.06
	SFT*	33.22	26.03	47.40	77.70	71.49	73.86	67.42	53.47	56.09	30.58
	SFT+DPO	32.34	25.81	46.73	76.22	71.09	72.94	66.87	53.96	56.26	31.01
	SFT+DPO*	28.81	25.84	45.96	74.30	**73.05**	73.04	65.00	52.78	54.79	30.84
	SFT+Ours	32.69	25.83	**47.97**	**78.64**	70.94	74.13	**69.05**	53.85	**57.00**	30.75
	SFT+Ours*	32.75	**26.17**	47.92	78.58	72.09	**74.60**	69.01	**54.03**	56.93	**32.01**

Table 6. Results of Qwen2-VL-7B training with DOCCI training set. BLEU-4, METEOR, CAPTURE and seven aspects of our proposed evaluation metrics are reported. * denotes metrics of the self-corrected captions. "Same-Domain" refers to the performance on DOCCI500 test set which has the same image and caption distribution as the training set. "Cross-Domain" denotes the performance on COCO-LN500 which has different distribution from the training set. Best results are highlighted in bold. Our proposed method outperforms baseline and DPO. Comparisons between "Same-Domain" in this table and Tab. 1 show that our proposed RefinedCaps dataset achieves comparable performance with the training set of the same domain (71.63 vs. 72.05 in Objects F1). However, the DOCCI training set performs worse in cross-domain scenario compared to the results in Tab. 2 (74.60 vs. 76.37 in Objects F1).

Since DOCCI500 test set is sampled from DOCCI, it can be referred to as a same-domain scenario. In contrast, COCO-LN500 represents a cross-domain scenario. The results in Tab. 6 can be compared with those in Tab. 1 and 2 to investigate the influence of training datasets. On DOCCI500 which is in the same domain as DOCCI training set, results of model trained on RefinedCaps are still comparable (71.63 vs. 72.05 in Objects F1, 57.67 vs. 58.00 in Attributes F1, 30.51 vs. 28.58 in Relations QA). However, on COCO-LN50 which is not the same domain as DOCCI training set, models trained on DOCCI performs much worse (74.60 vs. 76.37 in Objects F1, 56.93 vs. 57.56 in Attributes F1, 32.01 vs. 38.51 in Relations QA). The above in-domain and cross-domain analyses demonstrate that the generalization and adaptation ability of our proposed RefinedCaps dataset is stronger than DOCCI dataset in terms of supervised fine-tuning and self-correction training.

11.2. More Comparisons with Other Methods

To further demonstrate the effectiveness of our proposed method, more comparative experiments are conducted with different methods and the results are shown in Tab. 7. The first four lines are results reported in Tab. 1. DiscriTune is a reinforcement learning method introduced in [7], which utilizes CLIP [39] to produce reinforcement learning loss. Line 5 shows that this method fails to achieve satisfactory results. It may because CLIP truggles to effectively distinguish differences when dealing with very long captions. We also try to calculate the reward solely from the output of the first turn and put the results in line 6. It can boost performance, but fails to exceed the proposed two-step approach, demonstrating the necessity of self-correction. Additionally, we tested an extra baseline one where the SFT model is used to generate captions, which are then plugged into a $[x1, y1, x2]$ input mapped into a $y*$ output for a second phase of supervised finetuning. Results in lines 7-8 (SFT+SFT2) show that this setting fails to achieve better performance. In addition, results of only train the model as correcter (namely using the initial captions in RefinedCaps pipeline instead of the first-turn generated captions to calculate loss) are reported in the last line. This setting performs worse than the proposed approach with more training data (our proposed method only needs the final caption as GT).

Model	BLEU-4	METEOR	CAPTURE	O-F1	A-F1	Relations
None	29.39	16.59	57.96	66.47	52.65	17.57
SFT	40.92	22.04	62.05	69.50	55.50	27.65
SFT+DPO*	44.49	23.84	62.51	70.67	55.60	27.33
SFT+Ours*	**44.88**	**25.18**	**63.34**	**71.63**	**57.67**	**30.51**
SFT+DiscriTune	30.95	19.56	60.37	67.48	56.21	25.27
SFT+RL(1turn)	41.25	22.91	63.12	70.57	57.57	28.42
SFT+SFT2	41.76	22.61	61.97	69.43	55.18	29.39
SFT+SFT2*	40.19	22.09	62.03	70.00	56.02	29.79
SFT+Correction*	43.26	23.42	62.79	70.56	56.58	28.54

Table 7. Experimental results with more different methods. All experiments are based on Qwen2-VL-7B and DOCCI500.

11.3. Comparisons with more Models

We evaluate InternVL2-8B, ShareCaptioner, Gemini-1.5 and Claude-3.7 for image captioning and compare the results on DOCCI500 in Tab. 8. Closed-source models perform better than Qwen2-VL, but after training with proposed method, the baseline model can perform better.

GPT-4o (Atmosphere removed): A tennis player is wearing a white shirt and shorts, white wristbands and a headband, standing on the tennis court. The player is in a serving motion, with one arm extended holding a tennis racket. The racket's design is red, black, and white. The player's left foot is slightly lifted, and the body is leaning backward. The tennis court surface is divided into green and blue areas with white lines in the middle. In the background, part of the net can be seen, along with a few people sitting on the sidelines watching the match.

Human Refinement: A tennis player is wearing a white T-shirt and shorts, white wristbands and a white headband, standing on the tennis court. The player is in a serving motion, with one arm extended holding a tennis racket. The racket's design is red, black, and white. Both of the player's feet are slightly lifted, and the body is leaning backward. The tennis court surface is divided into green and blue areas with white lines in the middle.

GPT-4o (Atmosphere removed): Two people are standing indoors, wearing colorful umbrella hats. The person on the left is wearing a pink button-up shirt, with an arm draped over the shoulder of the person on the right, who is wearing a patterned dark gray T-shirt. The umbrella hats have multiple colors, including sections of red, green, yellow, and blue. Behind them is a kitchen scene, with white cabinets on the left, and a shelf on the right displaying various items such as jars, pots, and pans. On the counter in the foreground, a bowl can be seen.

Human Refinement: Two laughing people are standing indoors, wearing colorful umbrella hats. The person on the left is wearing a pink button-up shirt, with an arm draped around the waist of the person on the right, who is wearing a patterned dark gray T-shirt and black glasses. The umbrella hats have multiple colors, including sections of red, green, yellow, and blue. Behind them is a cluttered kitchen scene, with white cabinets on the left with some items on them, a glass kettle, and a black microwave underneath. The wooden shelves on the right display various items such as glass jars, black pots and pans, glass bottles, and white paper hanging on the edge. In the foreground on the counter, two inverted bowls can be seen. On the right, there is a white table with a yellow flower pot on it.

Figure 7. Visualization of human refinement samples in the RefinedCaps dataset. The red annotations represent description errors, and the green annotations represent the additional detail descriptions omitted by previous captions. Human annotators made meaningful improvements to enhance caption accuracy and completeness.

Model	BLEU-4	METEOR	CAPTURE	O-F1	A-F1	Relations
Qwen2-VL-7B	29.39	16.59	57.96	66.47	52.65	17.57
InternVL2-8B	28.70	16.92	58.30	66.69	53.41	16.85
ShareCaptioner	39.09	23.05	57.90	66.05	52.27	19.47
Gemini-1.5	24.70	16.25	60.34	68.17	55.69	28.90
Claude-3.7	41.36	20.17	61.02	69.48	55.57	28.78
Qwen2-VL-7B+SFT+Ours*	**44.88**	**25.18**	**63.34**	**71.63**	**57.67**	**30.51**

Table 8. Experimental results of more models on DOCCI500.

12. Visualization Examples

12.1. Annotated examples from RefinedCaps

To better illustrate the annotation process and the quality of the RefinedCaps dataset, we present sample cases with captions before and after human refinement in Fig. 7. These examples show that human annotators made meaningful improvements to enhance caption accuracy and completeness.

12.2. Comparisons among Methods

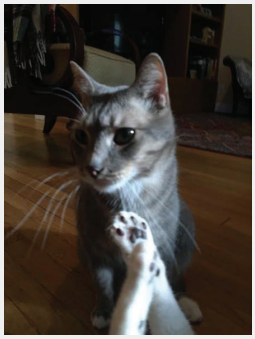
To directly show the difference among different post-training methods, we visualize the caption results of baseline, SFT, SFT+DPO and SFT+ours based on Qwen2-VL-7B. The qualitative results can be referred to in Fig. 8. It can be seen that captions of baseline models may miss some details, and our proposed SC-Captioner can describe more details while maintaining caption precision compared with SFT and DPO.

12.3. Self-Correction Effect

To intuitively display the self-correction effect of our proposed SC-Captioner, we visualize the initial and self-corrected captions in Fig. 9. It is shown that some incorrect descriptions are removed and some missing details are added during the self-correcting process.

13. Limitations and Future Work

The proposed framework requires ground-truth captions in the whole training process, adding to the data collection and annotation cost. Because the training process including generating annotations for two rounds, the training time is relatively long. Reinforcement learning methods that can self-explore without supervisions of ground-truth captions or even generating process may be directions for future research. Also, more metrics including reference-free ones like [14, 43] and LLM-assisted ones like [22, 54] can be tested and compared in the future.



Qwen2-VL: A gray cat is sitting on a wooden floor, reaching out with its paw towards the camera. The cat has a white paw and is positioned in a room with a bookshelf, a chair, and a rug in the background.

Qwen2-VL w/ SFT: A gray cat with white paws is sitting on a wooden floor, its front right paw raised towards the camera. The cat's eyes are open, and its ears are upright. In the foreground, the cat's paw is slightly blurred. In the background, on the left, there is a dark brown armchair with a patterned blanket draped over it, and on the right, there is a wooden bookshelf filled with books of various colors. A patterned rug can be seen on the floor.

Qwen2-VL w/ SFT+DPO: A gray cat is standing on a polished wooden floor, with its front paws raised. Its ears are upright, and its eyes are looking at the camera. In the foreground, a white object is partially visible, reaching towards the cat's paws. The cat's body is slightly turned to the left, with a dark chair covered with a textured blanket visible on the left side. In the background, there is a bookshelf filled with books of various sizes and a wooden floor lamp with a white shade. A patterned carpet is partially visible at the bottom of the bookshelf. The floor is a smooth, light-colored wood, and there is another dark chair in the background.

Qwen2-VL w/ SFT+Ours: A gray cat is sitting on a polished wooden floor. The cat's body is slightly tilted to the left, and its ears are perked up. In the foreground, the front legs of another animal, possibly a white dog, can be seen, with its paws raised towards the camera, as if waving. The cat's fur is gray with some white patches, and its eyes are wide open. In the blurred background, a wooden bookshelf is partially visible, with several books of different colors. On the left, a dark-colored armchair with a gray and white striped blanket hanging on the backrest is partially visible. The floor is brown wooden, and there is a patterned carpet in the room with a dark chair beside it. The walls are light-colored.

Qwen2-VL: The image shows a raccoon standing on a log in a body of water, surrounded by large green leaves. The raccoon appears to be exploring its environment, possibly searching for food or simply observing its surroundings. The background is filled with dense vegetation, creating a natural and lush setting.

Qwen2-VL w/ SFT: A small raccoon stands on a branch protruding from the murky water. Its fur is gray with darker stripes along the back and lighter fur on the face. The raccoon is positioned slightly to the left, surrounded by large green leaves, some of which are visible in the foreground. In the lower right corner, a colorful striped fabric or clothing can be seen.

Qwen2-VL w/ SFT+DPO: A raccoon with light brown and black fur stands on a partially submerged brown log in the shallow water. The raccoon's front paws are on the log, its body is slightly facing the camera, and its tail is visible, trailing behind it. Large green lotus leaves surround the raccoon, with a few leaves partially covering its head and shoulders. These leaves are distributed in the foreground and background, creating a lush green environment. In the lower right corner of the image, there is a colorful object resembling a piece of fabric, possibly part of a bag or clothing.

Qwen2-VL w/ SFT+Ours: A fluffy raccoon stands on a light-colored submerged log, surrounded by green leaves and green stems. The raccoon is slightly left of center in the image, facing the camera, with its paws partially on the log. The large green leaves and stems are spread across the frame, with some in the foreground and others in the background. These leaves are part of a plant that looks like taro, with broad and pointed leaves. On the right side of the image, part of a colorful garment with yellow, red, and green patterns can be seen, possibly a person's hand or clothing. The background is a dense area of leaves.

Figure 8. Additional qualitative results of baseline and three post-training approaches. The red annotations represent description errors, and the green annotations represent the additional detail descriptions omitted by previous captions. Our proposed method can reach more details while maintaining the precision of caption compared with baseline, SFT and DPO.

Qwen2-VL w/ SFT+Ours (Original): The mage shows a large monster truck, with a blue and red body and large black tires. The truck is prominently placed in the center of the indoor arena, surrounded by a blue barrier tape with white lettering. The truck's body features several prominent stripes, including a red horizontal stripe across the hood and another set of stripes on the sides. On the side of the truck, the number \"68\" can be seen, displayed on the fender. The truck is parked on a flat floor, and the ceiling is composed of multiple steel beams. In the background, part of the stadium's seating can be seen, including orange and black sections, with a few empty seats.

Qwen2-VL w/ SFT+Ours (Self-Correct): The mage shows a large colorful monster truck with a blue and red design, lifting its large front wheel high. The truck is prominently placed in the center of the indoor arena, surrounded by a blue barrier tape with white lettering. The truck's body features several prominent stripes, including a red horizontal stripe across the hood and another set of stripes on the sides. On the side of the truck, the number \"68\" can be seen, displayed on the fender. The truck is parked on a flat floor, and the ceiling is composed of multiple steel beams. In the background, part of the stadium's seating can be seen, including orange and black sections, with a few empty seats, with several white lights above. In the upper left corner of the image, a red and white American flag can be seen.

Qwen2-VL w/ SFT+Ours (Original): The picture shows a gray sports car, parked on a spacious indoor area with a gray floor, with a blue car behind it. The car is facing to the left, with its left side visible to the viewer. The car's roof is black, and the wheels are silver. In the background, a large window reveals an outdoor scene with green trees and a blue sky, and a tall grayish-white building on the left. In the distance, a large grayish-white bridge can be seen. The indoor space has large windows, with white light visible above the car.

Qwen2-VL w/ SFT+Ours (Self-Correct): The picture shows a gray sports car, which looks like a Porsche, parked on a spacious indoor area with a gray floor. The car is facing to the left, with its left side visible to the viewer. The car's roof is black, and the wheels are silver. In the background, a large transparent window reveals an outdoor scene with green trees and a blue sky, with some white clouds. On the left, there is a tall grayish-white building and a blue sign can be seen. In the distance, a large grayish-white bridge and a tall streetlight can be seen. The indoor space has large windows, with white light visible above the car.

Qwen2-VL w/ SFT+Ours (Original): A vintage blue Volkswagen Beetle is parked on the flat grass, with a wooden building in the background. The car is facing the left side of the image, and a white rack is installed on the roof. The vehicle is positioned slightly to the left of center. The building features a yellow exterior with horizontal wooden slats, and a transparent glass window on the right side of the image. The lower part of the building has a smooth concrete surface. In the foreground, the edge of a sidewalk can be seen. There are no visible texts or markings on the car or the building.

Qwen2-VL w/ SFT+Ours (Self-Correct): A vintage blue Volkswagen Beetle is parked on green grass, with a yellow wooden building in the background. The car is facing the left side of the image, and a white rack is installed on the roof. A chrome bumper is visible in front of the vehicle. The vehicle is positioned slightly to the left of center. The building features a yellow exterior with horizontal wooden slats, and a transparent glass window on the right side of the image, with a reflection surface. A clear tree reflection can be seen in the window. In the foreground, the gray edge of a concrete sidewalk can be seen. There are no visible texts or markings on the car or the building. There is a soft glow on the right side of the image.

Figure 9. Qualitative results of initial and self-corrected captions. The red annotations represent deleted descriptions, and the green annotations represent the added descriptions during self-correction. The self-correcting process can delete incorrect descriptions and add more details.