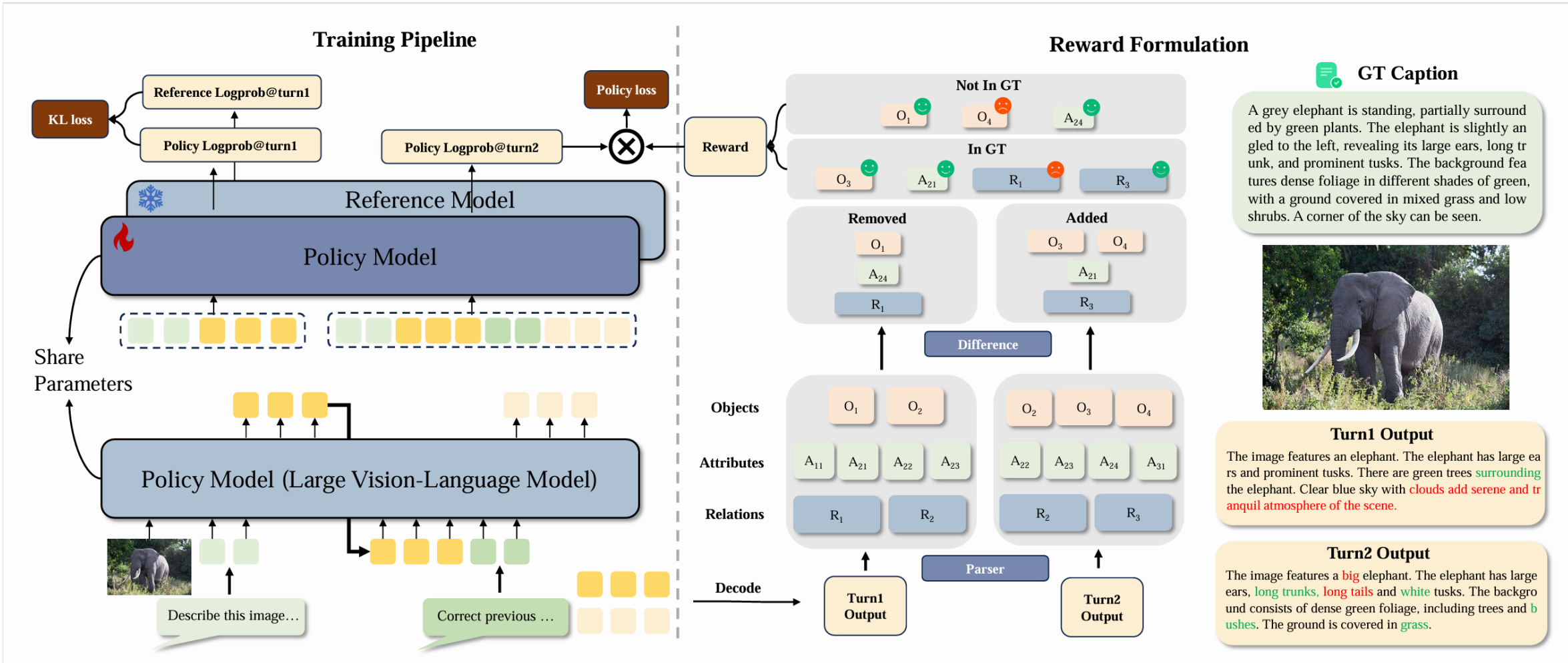
SC-Captioner: Improving Image Captioning with Self-Correction by Reinforcement Learning

Lin Zhang[1*], Xianfang Zeng[2♣], Kangcong Li[1], Gang Yu[2], Tao Chen[1,3†]
[1] College of Future Information Technology, Fudan University
[2] StepFun
[3] Shanghai Innovation Institute
22110720068@m.fudan.edu.cn, eetchen@fudan.edu.cn


Training Pipeline
Reference Logprob@turn1
KL loss
Policy Logprob@turn1
Policy loss
Policy Logprob@turn2
Reward
Reference Model
Policy Model
Share Parameters
Policy Model (Large Vision-Language Model)
Describe this image…
Correct previous …
Reward Formulation
Not In GT
In GT
Removed
Added
Difference
Objects
Attributes
Relations
Parser
Decode
Turn1 Output
Turn2 Output
GT Caption
A grey elephant is standing, partially surrounded by green plants. The elephant is slightly angled to the left, revealing its large ears, long trunk, and prominent tusks. The background features dense foliage in different shades of green, with a ground covered in mixed grass and low shrubs. A corner of the sky can be seen.
Turn1 Output
The image features an elephant. The elephant has large ears and prominent tusks. There are green trees surrounding the elephant. Clear blue sky with clouds add serene and tranquil atmosphere of the scene.
Turn2 Output
The image features a big elephant. The elephant has large ears, long trunks, long tails and white tusks. The background consists of dense green foliage, including trees and bushes. The ground is covered in grass.

GT Caption: "A rustic hanging bulbs fixture featuring thick ropes wrapped around a wooden beam, with bulbs hanging at different lengths. There is a vent top on the ceiling. The combination of natural materials and warm, soft lighting around the room are cozy."

Candidate Caption: "The hanging bulbs fixture showcases a unique interplay of elements, where thick ropes are meticulously draped and tied around a wood beam, creating a suspended frame. Vintage Edison-style bulbs hang from each rope end, shedding soft lighting. The chain suspends the entire structure from the ceiling. The ceiling has a vent. The warm and soft lighting create a comfortable atmosphere."

Objects

GT objects: bulbs, ropes, wooden beam, lighting, room, vent

Extra objects: door

CAPTURE: wood, beam, ropes, door, bulbs, lighting

Cand objects

Precision: 60.5
Recall: 52

Ours: ropes, wood beam, door, bulbs, lighting

Cand objects

Precision: 70
Recall: 58

Attributes

GT&Extra Attributes: ropes: thick, twisted; wooden beam: brown; lighting: warm, soft, pale; ...

CAPTURE: thick, twisted, brown, warm, pale, ropes, light, ...

Cand Attributes

Ours: thick, twisted, brown, warm, pale, ropes, light, ...

Cand Attributes

Relations

GT relations: (ropes) wrapped around (beam); (bulbs) hanging at (beam); (lighting) around (room); (vent) top on (ceiling)

Ours
Q1: Is the wooden beam tying around ropes? A1: Y.
Q2: Are the bulbs hanging at beam? A2: N.
Q3: Is the lighting around the room? A3: Y.
Q4: Is the vent top on the ceiling? A4: Y.

Cand answers: Y, N, N/A, Y

Score:70

CAPUTURE: (bulbs) hanging from (ropes); (bulbs) shedding (lighting); (chains) suspending from (ceiling); (ceiling) having (vent); (ropes) draped and tied around (beam); (beam) around (room)

Cand relations

Score: 38

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.

Base Model	Post-training	BLEU-4	METEOR	CAPTURE	Objects			Attributes			Relations
					Precision	Recall	F1	Precision	Recall	F1	QA
LLaVA-1.5-7B	Zero-shot	24.36	14.59	50.72	81.13	48.37	59.41	64.12	39.78	48.01	9.19
	Zero-shot*	23.09	14.08	50.47	**81.40**	48.25	59.25	64.70	40.10	48.39	9.23
	SFT	40.08	21.25	60.92	78.15	62.20	68.44	66.85	46.27	53.93	19.87
	SFT*	40.14	21.33	61.03	78.15	62.39	68.55	67.05	46.36	54.05	20.11
	SFT+DPO	42.38	22.90	61.30	76.64	62.88	68.26	66.22	46.83	54.15	20.64
	SFT+DPO*	42.67	23.47	61.09	75.31	64.75	68.89	65.64	46.29	53.61	21.73
	SFT+Ours	42.60	22.84	62.20	77.92	64.08	69.61	**67.81**	48.72	55.94	21.93
	SFT+Ours*	**43.04**	**23.88**	**62.29**	77.58	**66.05**	**70.30**	67.64	**50.70**	**57.10**	**22.69**
Qwen2-VL-7B	Zero-shot	29.39	16.59	57.96	**83.69**	56.79	66.47	69.96	43.27	52.65	17.57
	Zero-shot*	27.16	15.88	57.62	83.61	56.10	65.88	**70.36**	43.49	52.92	17.01
	SFT	40.92	22.04	62.05	79.99	62.68	69.50	68.85	47.57	55.50	27.65
	SFT*	38.93	21.34	62.02	80.46	62.19	69.32	69.18	47.91	55.78	27.93
	SFT+DPO	43.13	22.88	62.56	79.86	64.85	70.77	67.74	47.92	55.36	26.44
	SFT+DPO*	44.49	23.84	62.51	78.78	65.44	70.67	67.67	48.30	55.60	27.33
	SFT+Ours	43.70	23.35	63.00	80.48	64.48	70.74	69.68	48.83	56.61	29.06
	SFT+Ours*	**44.88**	**25.18**	**63.34**	80.20	**66.00**	**71.63**	69.54	**50.34**	**57.67**	**30.51**

Base Model	Post-training	BLEU-4	METEOR	CAPTURE	Objects			Attributes			Relations
					Precision	Recall	F1	Precision	Recall	F1	QA
LLaVA-1.5-7B	Zero-shot	**38.48**	20.60	44.75	**81.20**	59.02	67.56	58.61	37.20	42.34	14.38
	Zero-shot*	37.90	20.17	44.89	80.86	59.30	67.57	60.80	38.63	43.85	14.95
	SFT	33.81	26.29	46.62	77.09	71.16	73.45	66.65	50.57	54.25	28.59
	SFT*	33.60	26.37	46.61	77.15	71.55	73.72	**66.78**	50.70	54.43	27.82
	SFT+DPO	31.05	26.58	46.42	76.01	73.10	73.95	65.28	52.32	54.79	28.88
	SFT+DPO*	28.84	26.81	45.70	74.78	74.18	73.98	64.39	52.04	54.24	29.85
	SFT+Ours	32.72	26.76	46.57	76.89	73.49	74.64	66.18	51.90	54.87	32.37
	SFT+Ours*	31.60	**27.05**	**47.11**	76.43	**75.65**	**75.20**	65.79	**53.01**	**55.35**	**33.63**
Qwen2-VL-7B	Zero-shot	**39.57**	20.42	46.52	81.12	61.82	69.47	66.48	42.86	48.68	20.47
	Zero-shot*	38.99	19.58	46.61	**81.42**	61.71	69.52	66.78	42.97	48.81	21.16
	SFT	34.59	26.49	47.14	78.64	73.24	75.37	69.01	53.15	56.54	36.39
	SFT*	34.29	26.57	47.08	78.76	73.27	75.43	**69.02**	53.09	56.57	36.88
	SFT+DPO	31.78	26.88	46.71	77.51	73.38	74.86	67.68	52.28	55.35	36.23
	SFT+DPO*	30.34	26.77	46.44	77.01	74.23	75.14	67.28	52.65	55.79	37.21
	SFT+Ours	35.29	27.00	47.28	78.83	74.34	76.07	68.66	54.44	57.16	37.73
	SFT+Ours*	35.05	**27.34**	**47.51**	78.72	**75.01**	**76.37**	68.43	**55.11**	**57.56**	**38.51**