Cross-modal Identity Mapping: Minimizing Information Loss in Modality Conversion via Reinforcement Learning

Supplementary Material

7. Prompt Templates

We adopt simple and concise prompts to perform image captioning. For relation evaluation and hierarchical classification, we prompt Qwen3 [77] models to answer the given questions based on the candidate captions. The prompts are detailed as follows:

Prompt For Image Captioning
Caption this image as accurately as possible, without speculation. Describe what you see.

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>

Prompt For Species Classification
Caption: <Caption>
Check if the caption explicitly mentions any <Species>.
Answer exactly yes or no.

Prompt For Breed Classification
Caption: <Caption>
Check if the caption explicitly mentions the <Species> breed <Breed>.
Answer exactly yes or no.

Figure 6. **Prompts for image captioning, relation evaluation, and hierarchical classification.**

8. Additional Experiments

8.1. Comparisons with more Models

We evaluate Qwen2-VL-7B [72], ShareCaptioner [83], Gemini-1.5, and Claude-3.7 for image captioning and compare the results on the DOCCI500 [49] benchmark, as shown in Tab. 6. We observe that closed-source models (Gemini-1.5 and Claude-3.7) generally outperform the baseline Qwen2-VL-7B in CAPTURE, Object F1, Attribute F1, and Relation QA metrics. However, after applying our proposed reinforcement learning method CIM to the Qwen2-VL-7B SFT model, the model ('SFT + Ours') achieves the best performance across all metrics, surpassing both open-source and closed-source models.

Model	BLEU-4	METEOR	CAPTURE	Objects	Attributes	Relations
				F1	F1	QA
○ Qwen2-VL-7B [72]	29.39	16.59	57.96	66.47	52.65	17.57
○ ShareCaptioner [83]	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
• SFT + Ours	**45.16**	**24.88**	**64.31**	**73.87**	**58.68**	**36.32**

Table 6. **Performance comparison of open-source and closed-source models on the DOCCI500 [49] benchmark.** Best results are highlighted in bold. Our model achieves the best performance across all metrics, surpassing both open-source and closed-source models.

8.2. Comparisons with Larger Models

As shown in Tab. 7, reinforcement learning with CIM consistently boosts the performance of 7B/8B models, enabling them to surpass much larger models (72B/76B/78B) on both COCO-LN500 [54] and DOCCI500 [49]. Averaged over all backbones and benchmarks, our models achieve improvements of 2.3%, 3.6%, 2.0%, 4.3%, 5.0%, and 5.1% on BLEU-4, METEOR, CAPTURE, Objects F1, Attributes F1, and Relation QA compared to their larger counterparts. These results demonstrate that our CIM-based reinforcement learning framework has substantially unlocked the potential of 7B/8B models, enabling them to generate more detailed and precise captions than their much larger counterparts.

8.3. Training on the DOCCI Training Set

We also use the training set of DOCCI [49] which consists of 9.7K image-caption pairs as the training set for supervised fine-tuning and reinforcement learning of Qwen2-VL-7B [72]. Evaluation results on the COCO-LN500 [54] and DOCCI500 [49] benchmarks are reported in Tab. 8. Our proposed method consistently outperforms both the SFT model and SC-Captioner [83] by a substantial margin. For instance, compared with SC-Captioner, our models achieve improvements of 1.2% in Attributes F1 on COCO-LN500 and 1.3% on DOCCI500. Similarly, it achieves improvements of 4.8% in Relations QA on COCO-LN500 and 7.3% on DOCCI500. These consistent gains across both datasets highlight the robustness and generalizability of our reinforcement learning approach.

Base Model	Benchmark	Method	BLEU-4	METEOR	CAPTURE	Objects			Attributes			Relations
						Precision	Recall	F1	Precision	Recall	F1	QA
Qwen2-VL [72]	COCO-LN500	○ 7B-Base	39.57	20.42	46.52	81.12	61.82	69.47	66.48	42.86	48.68	20.47
		○ 72B-Base	39.37	19.77	47.20	**82.52**	62.89	70.72	69.36	42.78	49.38	27.46
		● 7B-Ours	**39.63**	**27.06**	**48.64**	80.90	**72.23**	**75.80**	**72.45**	**54.08**	**58.22**	**38.71**
	DOCCI500	○ 7B-Base	29.39	16.59	57.96	83.69	56.79	66.47	69.96	43.27	52.65	17.57
		○ 72B-Base	30.79	17.57	59.56	**84.02**	60.21	69.14	71.63	44.31	53.92	27.21
		● 7B-Ours	**37.21**	**21.53**	**63.12**	81.99	**64.61**	**71.43**	**74.49**	**50.19**	**59.18**	**32.12**
Qwen2.5-VL [4]	COCO-LN500	○ 7B-Base	29.11	14.58	44.12	**82.35**	55.72	65.37	66.30	39.90	46.25	23.76
		○ 72B-Base	**40.21**	22.57	48.06	81.21	66.37	72.42	68.46	48.21	52.97	32.13
		● 7B-Ours	37.15	**26.81**	**48.93**	79.28	**75.91**	**77.59**	**72.49**	**54.46**	**58.51**	**44.15**
	DOCCI500	○ 7B-Base	22.68	14.67	55.89	**84.64**	54.96	65.06	72.15	42.13	52.27	24.35
		○ 72B-Base	33.72	19.01	61.09	82.17	62.92	70.35	72.37	46.18	55.46	30.71
		● 7B-Ours	**39.80**	**22.72**	**63.46**	79.28	**66.97**	**71.88**	**73.52**	**50.45**	**59.08**	**34.70**
InternVL2 [8]	COCO-LN500	○ 8B-Base	31.44	21.28	45.86	80.46	65.20	71.05	70.63	47.40	52.73	26.65
		○ 76B-Base	37.31	22.29	47.64	80.57	67.25	72.66	70.69	46.90	54.50	31.60
		● 8B-Ours	**41.89**	**25.85**	**49.06**	**82.05**	**69.40**	**74.57**	**73.50**	**52.59**	**57.49**	**35.54**
	DOCCI500	○ 8B-Base	31.49	17.97	58.83	81.49	59.54	67.72	70.84	44.24	53.51	22.65
		○ 76B-Base	**38.69**	**21.21**	**61.09**	81.42	**64.77**	**71.19**	70.90	46.80	55.57	**29.67**
		● 8B-Ours	32.01	19.49	60.82	**82.16**	60.35	68.66	**74.16**	**47.51**	**56.98**	26.36
InternVL2.5 [7, 74]	COCO-LN500	○ 8B-Base	38.04	18.54	47.09	**82.42**	64.02	71.42	**72.00**	46.03	52.50	27.29
		○ 78B-Base	**41.22**	20.79	48.25	81.38	64.59	71.42	71.25	46.84	52.85	28.92
		● 8B-Ours	39.27	**22.85**	**48.28**	80.31	**70.42**	**74.53**	71.85	**51.71**	**56.07**	**36.96**
	DOCCI500	○ 8B-Base	24.23	15.29	58.64	84.11	58.07	67.62	73.48	44.79	54.69	24.63
		○ 78B-Base	23.60	15.79	59.37	**85.78**	59.12	68.80	**75.12**	45.50	55.74	26.92
		● 8B-Ours	**30.89**	**17.84**	**60.75**	82.17	**62.28**	**69.99**	72.96	**46.64**	**56.07**	**28.90**
InternVL3 [86]	COCO-LN500	○ 8B-Base	33.21	16.16	47.88	**82.51**	63.31	71.00	71.14	43.97	50.66	26.44
		○ 78B-Base	40.60	19.50	48.48	82.46	65.02	72.07	71.76	46.22	52.50	29.61
		● 8B-Ours	**40.64**	**25.64**	**48.90**	80.33	**73.33**	**76.14**	**73.33**	**54.54**	**58.70**	**38.67**
	DOCCI500	○ 8B-Base	12.57	11.66	56.95	85.98	55.39	66.08	74.75	43.19	53.72	25.11
		○ 78B-Base	18.32	13.89	58.52	**86.18**	58.23	68.30	75.69	44.33	55.01	27.85
		● 8B-Ours	**28.25**	**18.54**	**62.18**	82.73	**62.86**	**70.47**	**76.22**	**49.58**	**59.26**	**30.39**

Table 7. **Performance of Reinforcement Learning with CIM on Base Models over COCO-LN500 [54] and DOCCI500 [49].** Best results are highlighted in bold. The results indicate that post-training the base model with our CIM enhances its ability to generate high-quality captions, and even 7B/8B models can outperform much larger baselines.

In the COCO-LN500 [54] setting, the ground-truth captions are generally shorter, which leads to relatively low precision scores. This occurs because shorter captions often omit certain objects and details present in the images, thereby penalizing the precision metric even when these elements are correctly identified by the model. Similarly, the highest BLEU-4 score observed for the base model can be attributed to this characteristic of the dataset.

Base Model	Benchmark	Method	BLEU-4	METEOR	CAPTURE	Objects			Attributes			Relations
						Precision	Recall	F1	Precision	Recall	F1	QA
Qwen2-VL-7B [72]	COCO-LN500	○ Base	**39.57**	20.42	46.52	**81.12**	61.82	69.47	66.48	42.86	48.68	20.47
		○ SFT	34.93	25.97	47.63	78.03	70.42	73.43	67.91	53.51	56.31	30.06
		○ SC-Captioner [83]	32.75	26.17	47.92	78.58	72.09	74.60	69.01	54.03	56.93	32.01
		• SFT + Ours	32.34	**26.50**	**48.36**	78.76	**72.93**	**75.13**	**69.72**	**55.23**	**58.10**	**36.76**
	DOCCI500	○ Base	29.39	16.59	57.96	**83.69**	56.79	66.47	69.96	43.27	52.65	17.57
		○ SFT	40.29	25.44	62.53	78.01	65.30	70.31	67.13	49.33	56.08	25.43
		○ SC-Captioner [83]	41.95	26.38	63.83	78.85	67.59	72.05	69.77	50.64	58.00	28.58
		• SFT + Ours	**42.62**	**27.13**	**64.48**	80.03	**68.59**	**73.10**	**71.83**	**51.58**	**59.30**	**35.91**

Table 8. **Performance of Reinforcement Learning with CIM on the SFT Model, trained on the DOCCI [49] training set, over COCO-LN500 [54] and DOCCI500 [49].** Best results are highlighted in bold. Our model outperforms both the SFT model and SC-Captioner [83].

9. Visualization Examples

9.1. Comparisons with the Base Model

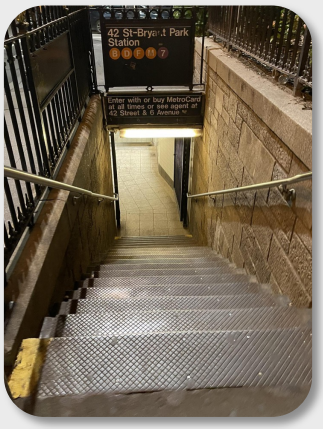
To directly illustrate the effect of reinforcement learning with CIM on the base model, we present qualitative examples comparing the original base model and the base model trained with CIM, as shown in Fig. 7. We mark correct descriptions in green for clarity. As can be observed, captions generated by the base model often overlook fine-grained details, whereas our method is able to generate more comprehensive descriptions, demonstrating improved descriptive capability.

9.2. Comparisons with the SFT Model

To further validate the effectiveness of our approach, we provide a qualitative comparison among the base model, the SFT model, and the SFT model trained with reinforcement learning using CIM, as shown in Fig. 8. Correct descriptions are marked in green and incorrect descriptions in red. It can be observed that captions generated by the base model often miss fine-grained details. In contrast, our proposed method can generate more detailed descriptions while maintaining caption precision compared with SFT model.

Qwen2.5-VL-7B: A white handprint is painted on a red wall. The background wall has a rough, speckled surface, and the handprint is centered in the image.

Qwen2.5-VL-7B w/ Ours: A white handprint on a red wall. The handprint is in the center of the image, with the thumb on the left and the pinky on the right. The wall has a textured surface and a few small marks and imperfections scattered around the handprint. There is a horizontal line running across the wall below the handprint.



Qwen2.5-VL-7B: Stairs leading to the 42nd Street-Bryant Park Station, with signage indicating the train lines (B, D, F, M, 7) and a note about entering with or buying a MetroCard.

Qwen2.5-VL-7B w/ Ours: Stairs leading down to a subway station. Above the stairs, there is a sign that reads "42 St-Bryant Park Station" with orange circles containing the letters B, D, F, M, and 7. Below the sign, there is another sign that says "Enter with or buy MetroCard at all times or see agent at 42 Street & 6 Avenue." On both the left and right sides of the stairs, there is a concrete wall with a metal railing. At the top of the stairs, there is a black metal railing with ornate designs. In the background, there is a street with a car parked on it.

Qwen2.5-VL-7B: Stacked blocks with circular patterns on a playground railing.

Qwen2.5-VL-7B w/ Ours: Three stacks of large, beige cylinders with a black circle in the center, stacked on top of each other, placed on a metal table with holes in it. Behind the cylinders, there is a playground with a child in a green shirt running on a black surface. In the background, there is a grassy area with trees and a house on the left side. A yellow slide is partially visible on the right side of the image. The cylinders are placed on a beige railing.

Figure 7. **Qualitative comparison between the base model and the base model trained with reinforcement learning using CIM.** Green annotations indicate correct descriptions. Our method enhances the base model's ability to produce more detailed captions.

Qwen2-VL-7B: The image shows a wooden surface divided into two sections. The top section is dark brown, while the bottom section is light brown. On the bottom section, there are four red and blue crosses painted on the wooden planks. The crosses are evenly spaced and have a distinct design, with a blue cross in the center and a red cross on each side.

Qwen2-VL-7B w/ SFT: The image shows a wooden surface divided into two horizontal sections. The upper section is dark gray, while the lower section is light brown. On the lower section, there are four evenly spaced blue and red crosses. The blue crosses are on the left and right sides, and the red crosses are in the middle. The crosses are painted in bright colors, contrasting with the wooden background.

Qwen2-VL-7B w/ SFT + Ours: The image shows a wooden fence with two distinct sections of planks. The top section is dark brown, while the bottom section is light brown. On the light brown section, there are four evenly spaced blue and red crosses. From left to right, the first cross is blue, the second is red, the third is blue, and the fourth is red. Painted in bright colors, the crosses contrast with the wooden background.

Qwen2-VL-7B: The image shows five bananas arranged in a circular pattern on a wooden surface. The bananas are yellow with some brown spots, indicating they are ripe. The arrangement creates a circular shape with the bananas forming the outline.

Qwen2-VL-7B w/ SFT: Five yellow bananas are arranged in a circular pattern on a wooden surface. The bananas are evenly spaced, forming a complete circle. The wooden surface has a natural grain pattern, with visible knots and wood grain. The bananas are placed on the surface, with the stems pointing outward.

Qwen2-VL-7B w/ SFT + Ours: Six yellow bananas are arranged in a circular pattern on a wooden table. The bananas are evenly spaced, creating a complete circle. The wooden table has a rustic appearance, with visible grain and knots. The bananas are placed on the table, each with its stem pointing toward the tip of the next.

Figure 8. **Qualitative comparison among the base model, the SFT model, and the SFT model trained with reinforcement learning using CIM.** Green annotations indicate correct descriptions, while red annotations indicate incorrect descriptions. Compared with the base model and the SFT model, our method produces more detailed captions while maintaining the precision of caption.