Supplementary Materials for "CCCaption: Dual-Reward Reinforcement Learning for Complete and Correct Image Captioning Supplementary Materials"

Zhijiang Tang[1,2*], Linhua Wang[3‡], Jiaxin Qi[2*], Weihao Jiang[3],
Peng Hou[3], Anxiang Zeng[3†], Jianqiang Huang[1,2†]
[1]Computer Network Information Center, Chinese Academy of Sciences, Beijing, China
[2]Hangzhou Institute for Advanced Study, University of Chinese Academy of Sciences, Zhejiang, China
[3]LLM Team, Shopee Pte. Ltd., Shanghai, China
tangzhijiang24@mails.ucas.ac.cn, {jxqi, jqhuang}@cnic.cn,
{wanglinhua, weihao.jiang, peng.hou, anxiang.zeng}@shopee.com

The supplementary material provides comprehensive details to complement the main paper. **Regarding the Method**, we elaborate on the Group Relative Policy Optimization (GRPO) objective and the implementation of the Dynamic Query Sampling strategy, alongside formal definitions of the evaluation frameworks. **Regarding Implementations**, we provide detailed descriptions of the dataset characteristics and prompt designs. **Regarding Experiments**, we present the complete quantitative results from the main paper, supplemented by qualitative case studies and an extended discussion on evaluation robustness and computational overhead.

Method

Reinforcement Learning

Group Relative Policy Optimization (GRPO) [7] is a relatively advanced reinforcement learning algorithm. Given an image captioning dataset $\mathcal{D}=\{(\mathbf{x},\mathbf{y})\}$, where $\mathbf{x}$ denotes images and $\mathbf{y}$ denotes captions, the GRPO objective can be written as maximizing the expected reward:

$$\mathcal{J}_{\text{RL}}(\theta) = \frac{1}{G}\sum_{i=1}^{G}\frac{1}{|\mathbf{y}_i|}\sum_{t=1}^{|\mathbf{y}_i|}\Bigg\{ \min\left[\frac{\pi_\theta(y_{i,t}|y_{i,<t},\mathbf{x})}{\pi_{\theta_{\text{old}}}(y_{i,t}|y_{i,<t},\mathbf{x})}\hat{A}_{i,t}, \text{clip}\left(\frac{\pi_\theta(y_{i,t}|y_{i,<t},\mathbf{x})}{\pi_{\theta_{\text{old}}}(y_{i,t}|y_{i,<t},\mathbf{x})}, 1-\varepsilon, 1+\varepsilon\right)\hat{A}_{i,t}\right] - \beta\mathbf{D}_{\text{KL}}\left[\pi_\theta||\pi_{\text{ref}}\right]\Bigg\}$$

where θ denotes the trainable parameters of the policy model π_θ. $\mathbf{y}_i$ is the i-th response sampled from a group of size G. The term $y_{i,t}$ denotes the token at step t, and $y_{i,<t}$ represents the preceding tokens. $\hat{A}_{i,t}$ is the advantage estimate for the i-th response, calculated via group-relative normalization (i.e., $\hat{A}_{i,t}=\widetilde{r}_i=\frac{r_i-\text{mean}(\mathbf{r})}{\text{std}(\boldsymbol{r})}$), where $\boldsymbol{r}=\{r_i\}_{i=1}^{G}$ is the reward vector over the group. ε is the clipping hyperparameter, and β controls the weight of the KL divergence penalty between the current policy π_θ and the reference policy π_{ref}.

Dynamic Query Sampling. Building upon the concept introduced in the main paper, we implement the dynamic query sampling strategy. This approach is designed to enhance training efficiency by prioritizing samples that provide the greatest informational gain. Specifically, we sample queries where the model's rollout accuracy exhibits high variance, as these are the queries most likely to yield non-negligible gradients. This focused sampling allows the optimization process to concentrate on the most challenging or informative instances. The set of dynamically sampled queries, $\tilde{\mathcal{Q}}$, is drawn from the candidate set $\mathcal{Q}_{\text{c}}$ using a Bernoulli trial [19]. The probability of a query $\mathbf{q}$ being selected is governed by its normalized contribution $\tilde{c}(\mathbf{q})$.

$$\tilde{\mathcal{Q}} = \left\{\, \mathbf{q}\in\mathcal{Q}_{\text{c}} \,\middle|\, u_{\mathbf{q}}=1,\ u_{\mathbf{q}}\sim\text{Bernoulli}(\tilde{c}(\mathbf{q}))\right\},$$

where $u_{\mathbf{q}}$ is the sampling indicator. The normalized contribution $\tilde{c}(\mathbf{q})=c(\mathbf{q})/\sum_{\mathbf{o}\in\mathcal{Q}_{\text{c}}}c(\mathbf{o})\in[0,1]$. And the contribution $c(\mathbf{q})$ is initialized based on the query's initial failure rate and updated iteratively to reflect the instability of the current policy:

$$c(\mathbf{q})^{(0)} = 1-\frac{1}{m}\sum_{i=1}^{m}\mathbb{I}(M_J(\mathbf{x},\mathbf{q}))$$

where m is the number of initial sampling, $\mathbf{x}$ is the input

[1]Equal contribution. Work done during Z. Tang's internship at Shopee.
[3]Project lead.
[2]Corresponding authors.

Prompt for Captioning Model

Describe this image in details.

Figure 1. Prompt for captioning model. To ensure consistency, this prompt is used by all captioning models during both training and evaluation.

Prompt for the Completeness Judge Model

You are an AI assistant that answers multiple-choice questions only based on an image caption.
You must carefully read the caption and select the most appropriate answer from the given options (A, B, C, or D).
Your output must only contain one letter: A, B, C, or D.
Do not include explanations or extra text.

Example
Image Caption:
A group of students is sitting in a classroom, listening to a teacher who is writing on the whiteboard.
Question:
What are the students most likely doing?
A. Eating lunch
B. Playing soccer
C. Attending a lesson
D. Taking a nap

Your Response:
C

Image Caption: **{Caption}**
Question: **{Query}**

Figure 2. Prompt for completeness judge model. The text highlighted in red denotes placeholders that are substituted with the specific caption and query during implementation.

image, and $\mathbb{I}(\cdot)$ is an indicator function that returns 1 when the answer is correct, determined by the MLLM judge M_J.

At each epoch i, the contribution is updated by incorporating the variance of the rollout accuracy for that query, promoting those queries where the model's performance is highly inconsistent:

$$c(\mathbf{q})^{(i)} = c(\mathbf{q})^{(i-1)} + \lambda \cdot \frac{1}{G} \sum_{i=1}^{G} (r_{q,i} - \text{mean}(\boldsymbol{r}_\mathbf{q}))^2$$

where $\boldsymbol{r}_\mathbf{q} = \{r_{q,i}\}_{i=1}^{G}$ is the query reward vector over the group, and λ is a hyperparameter balancing the influence of the variance term.

Evaluation Frameworks

Prism Framework [18] evaluates caption quality based on the accuracy with which a judge model answers questions derived from the generated caption. Captions that answer more visual questions indicate that they cover more details of the images. The calculation of the prism score is as follows:

$$S_{\text{prism}}(\theta) = \frac{1}{|\mathcal{D}|} \sum_{\mathbf{x},\mathbf{q} \in \mathcal{D}} \mathbb{I}\big(M_J(M_\theta(\mathbf{x}), \mathbf{q})\big),$$

Prompt for the Correctness Judge Model

You are shown an image and a short subtitle.
Your task:
1. Split the subtitle into simple parts (objects, attributes, or actions).
2. Compare each part with the image.
3. Give one realism score (1–5).
Higher = more accurate, less hallucination.

Examples:

Image: a brown dog running on grass
Subtitle: "A white cat sitting on the floor"
→ Score: 1
Subtitle: "a black dog running on grass"
→ Score: 2
Subtitle: "A small brown dog on the grass"
→ Score: 3
Subtitle: "A small brown dog ran freely across the grass."
→ Score: 1
Subtitle: "a brown dog running on grass "
→ Score: 5

Now, for the next input, output only one number (1–5) as the score.
Subtitle: **{Subtitle}**

Figure 3. Prompt for correctness judge model. The text highlighted in red denotes a placeholder that is substituted with the sub-caption query during implementation.

where $\mathcal{D} = \{(\mathbf{x}, \mathbf{q})\}$ is a evaluation dataset, which $\mathbf{x}$ denotes a image and $\mathbf{q}$ denotes the relative visual question. M_θ is the captioning model and M_J is a frozen judge model that answers $\mathbf{q}$ based on the generated caption $M_\theta(\mathbf{x})$.

Hallucinations Framework [10] splits the image caption into atomic captions, then uses MLLMs to detect whether these atomic captions contain hallucination. The calculation of the hallucination score is as follows:

$$S_{\text{hall}}(\theta) = \frac{1}{|\mathcal{D}|} \sum_{\mathbf{x} \in \mathcal{D}} M_J(M_\theta(\mathbf{x}), \mathbf{x}),$$

where the judge model M_J give hallucination score between the generated caption $M_\theta(\mathbf{x})$ and the image $\mathbf{x}$.

CapArena Framework [4] establishes an arena-style evaluation platform comprising 600 pairwise caption battles, with the outcome of each comparison determined by a MLLM judge. The CapArena score is calculated as the mean score obtained across the evaluation dataset $\mathcal{D}$:

$$S_{\text{arena}}(\theta, \theta_{\text{ref}}) = \frac{1}{|\mathcal{D}|} \sum_{\mathbf{x} \in \mathcal{D}} \mathbb{J}(M_J(M_\theta, M_{\theta_{\text{ref}}}))$$

where M_θ denotes the candidate captioning model being evaluated, and $M_{\theta_{\text{ref}}}$ is a reference captioning model. The MLLM judge, M_J, performs the pairwise comparison, and the function $\mathbb{J}$ assigns the resulting score to the candidate model M_θ: $+1$ (superior), 0 (equivalent), or -1 (inferior) performance relative to $M_{\theta_{\text{ref}}}$'s caption for the image $\mathbf{x}$.

Prompt for Multi MLLMs Production

Your task is to generate five new and challenging multiple-choice questions (with answers) about the object based on the provided image.
Important constraints:
- The generated questions **must not duplicate or paraphrase** any existing questions in the provided QA list.
- You should analyze the provided QA array `{QA}` and ensure that all new questions are **novel**, **non-overlapping**, and explore **different aspects** or **details** of the image.
- Avoid repeating topics, phrasing, or answers from the existing QA.
- Focus on fine-grained or contextual details to make the questions challenging.
Output format:
You must output strictly valid JSON — a list of exactly 5 QA objects.
Each QA object must contain:
  - "question": the question text
  - "options": a dictionary with keys "A", "B", "C", "D" for the multiple-choice options
  - "answer": the correct answer option (e.g., "D")
Example format:

```
"'json
  [
  {
    "question": "Which method achieves the highest accuracy (Acc) on the FF++ (HQ) dataset?",
    "options": {
    "A": "Method 'a'",
    "B": "Method 'b'",
    "C": "Method 'c'",
    "D": "Ours"
    },
    "answer": "D"
  },
  ...
  ]
"'
```

Output requirements:
- Five Exactly and Challenging QA objects.
- Output must be valid JSON (no extra text, no explanations, no markdown).
- All questions and answers must be based on the provided image and must be new relative to the existing QA list.

Existing QA list: **{QA}**

Figure 4. Prompt for multi-MLLM production is utilized to generate the training set CCaption-44k, which is designed for comprehensive coverage of the key image information. The text highlighted in red denotes a placeholder that is substituted with the specific Query list during implementation.

Implementations

Evaluation Framework

Below, we provide a detailed introduction to the datasets.

Prism Framework [18, 22]. The primary datasets:

- ChartQA [14]: A dataset focused on evaluating chart comprehension that combines human-written questions with machine-generated summaries. It rigorously tests logical and linguistic understanding across various chart types (e.g., bar, line, and pie charts).
- CharXiv [21]: A challenging benchmark featuring real-world scientific charts extracted from arXiv research papers. It is explicitly designed to test MLLMs' ability to interpret complex, domain-specific visualizations that require expert-level knowledge.
- InfoVQA [15]: A visual question-answering dataset centered on infographics and document understanding. It features high-resolution images with diverse textual and visual elements, designed to test a model's multimodal capabilities for locating, reading, and reasoning.
- MathVerse [25]: A holistic benchmark for visual mathematical reasoning (Verse) that covers multiple subjects

MLLM	ChartQA [14]	CharXiv [21]	InfoVQA [15]	Verse [25]	MMB [12]	MMMUPro [24]	OCR [13]	COCO [8]	WM2Pro [17]	Avg.
LLaVa-V1.6-34B [11]	52.60	26.33	34.46	27.45	71.61	33.33	34.95	37.08	32.67	38.94
Qwen2.5-VL-3B [1]	69.72	31.66	51.81	29.53	72.58	36.97	44.01	38.80	36.88	45.77
Qwen2.5-VL-72B [1]	76.00	35.74	**59.40**	33.16	78.03	**46.26**	46.93	46.80	40.05	51.37
CapRL-3B [22]	73.92	35.11	55.94	33.16	79.16	41.01	50.16	52.53	38.64	51.07
InternVL3.5-38b [16]	74.84	**37.30**	**59.40**	34.91	77.04	41.62	50.81	47.28	38.82	51.34
Qwen3-VL-8B [20]	**76.16**	36.05	53.92	34.65	80.97	43.43	48.22	54.39	**40.29**	52.01
Qwen3-VL-32B [20]	75.72	35.74	51.94	**35.82**	**81.34**	43.64	49.19	**56.82**	40.14	52.26
Qwen3-VL-2B [20]	71.00	31.66	51.78	33.23	79.44	44.24	50.16	50.41	38.58	50.05
CCCaption-2B (Ours)	75.12	36.99	55.18	35.11	79.28	44.65	**55.34**	53.70	39.83	**52.80**

Table 1. Full results of the Prism evaluation [18] across different captioning models. Bold numbers indicate the best performance, underlined numbers indicate the runner-up.

MLLM	AI2D [9]	ChartQA [14]	Hallusion [6]	MMB [12]	MME [2]	MMMU [23]	MMStar [3]	OCR [13]	WM2Pro [17]	Avg.
LLaVa-V1.6-34B [11]	51.03	60.55	74.52	70.22	68.95	74.27	64.01	77.19	64.94	67.30
Qwen2.5-VL-3B [1]	48.71	61.73	74.76	68.96	65.81	75.77	63.46	78.09	72.66	67.77
Qwen2.5-VL-72B [1]	50.95	64.96	75.03	69.82	65.85	78.97	65.08	77.69	**75.12**	69.27
CapRL-3B [22]	48.62	63.86	73.03	67.41	62.65	76.01	61.98	76.19	72.44	66.91
InternVL3.5-38b [16]	50.86	64.11	75.93	**71.10**	**68.15**	78.42	65.32	82.72	**75.12**	70.19
Qwen3-VL-8B [20]	51.78	63.09	76.50	70.37	64.77	**79.12**	64.19	81.94	75.05	69.65
Qwen3-VL-32B [20]	**52.38**	**66.82**	75.45	69.20	64.38	79.03	64.37	82.35	74.87	69.87
Qwen3-VL-2B [20]	51.51	64.53	**77.51**	70.91	66.78	78.85	64.83	82.32	74.45	70.19
CCCaption-2B (Ours)	51.88	66.72	76.72	70.91	66.16	79.07	**65.50**	**83.09**	75.10	**70.57**

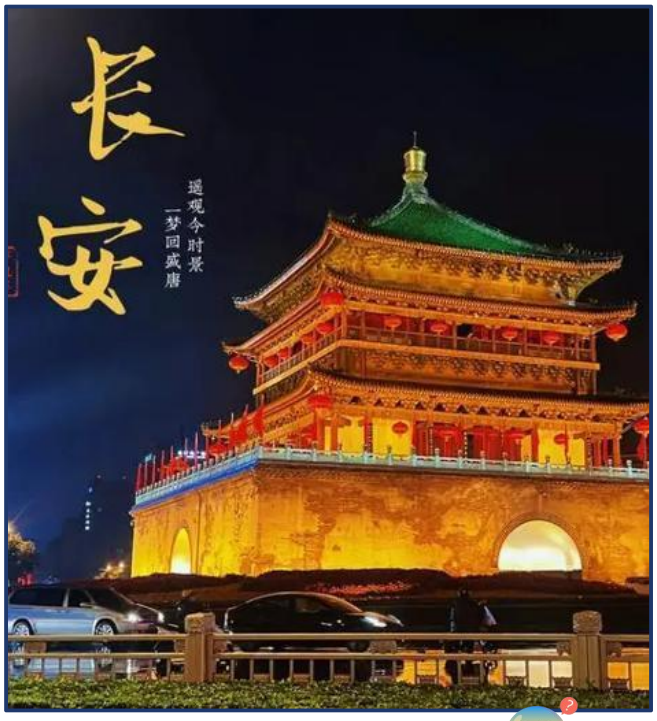
Table 2. Full results of the Hallucinations evaluation [10] across different captioning models. Bold numbers indicate the best performance, underlined numbers indicate the runner-up.

such as plane geometry and functions. It employs a unique evaluation strategy by removing textual redundancy to assess the model's ability strictly.

- MMBench [12]: A comprehensive multimodal understanding benchmark (MMB) that spans diverse task types, including logical reasoning and coarse perception. It utilizes a robust circular evaluation strategy (CircularEval) to mitigate random guessing.
- MMMUPro [24]: An advanced benchmark for expert-level interdisciplinary understanding, acting as a stricter evolution of MMMU. It covers a wide range of professional subject areas and filters for questions that require rigorous visual processing.
- OCRBench [13]: A dedicated benchmark for Optical Character Recognition (OCR) that aggregates 29 diverse datasets. It evaluates text recognition, document-based Q&A, and key information extraction, providing a unified score to assess a model's text-centric visual perception.
- RefCOCO [8]: A foundational dataset designed to evaluate object localization and referring expression comprehension. It tests an MLLM's ability to ground natural-language descriptions into specific bounding boxes within an image, assessing spatial awareness.
- WeMath2.0-Pro [17]: A specialized dataset specifically crafted for advanced visual mathematical reasoning (WM2Pro). It upgrades previous iterations by focusing on more challenging, multi-step mathematical problems that require a deeper integration of algebraic logic.

Hallucinations Framework [5, 10]. We supplement the following datasets to detect hallucinations:

- AI2D [9]: A dataset comprising over 4,900 diagrams from grade-school science textbooks. It is richly annotated with varying semantic layers and linked questions, requiring models to understand the interplay between textual labels and arrows.
- HallusionBench [6]: An advanced diagnostic benchmark specifically curated to provoke and analyze model hallucinations. It emphasizes visual illusions and nuanced image consistency checks.
- MME [2]: A comprehensive evaluation suite covering 14 distinct subtasks of perception and cognition across multimodal scenarios. It is particularly effective for halluci-



CCCaption-2B

.....he tower is adorned with red lanterns, which are hung along the eaves and at the top, adding a festive and vibrant touch to the scene. The base of the tower is a large, arched gateway, and the entire structure is set against a dark, night sky.

*- **The Setting:** The tower is situated in an urban environment, with modern vehicles visible in the foreground. A silver car and a black car are parked or moving along a road that runs in front of the tower. A low, railing separates the road from the base of the tower. The surrounding area is dark, with only the tower and the vehicles being illuminated.*

*- **Text and Composition:** The image is overlaid with large, golden Chinese characters on the left side. The character "长安" (Cháng'ān), which means "Xian" or "Xi'an," is written in a large, stylized font. To the right of the character, smaller text reads "遥观今时景 一梦回盛唐" (Yáo guān jīn shí jǐng yī mèng huí shèng táng).....*

Qwen3-VL-2B

Nitpicking

好的，这是一张关于中国著名历史古迹的夜景照片。

这张照片的主体是一座宏伟的古代楼阁，其建筑风格具有鲜明的中国唐代特色，很可能就是西安钟楼。

- **建筑主体**：照片的中心是一座多层的古代楼阁，其结构为典型的中国古代楼阁式建筑。它拥有层层叠叠的飞檐，屋檐上装饰着精致的斗拱和雕刻，屋顶覆盖着绿色的琉璃瓦。建筑的每一层都悬挂着红色的灯笼，这些灯笼在夜色中显得格外醒目。楼阁的底部是一个巨大的拱形门洞，门洞内透出温暖的光，与楼阁的金色灯光相映成趣。整个建筑被温暖的黄色灯光照亮，显得庄重而辉煌。

- **背景与环境**：[illegible]的背景是漆黑的夜空，与灯火通明的楼阁形成鲜明对比，突显了其历史的厚重感。在建筑的前方，[illegible]道路，路上有几辆汽车，包括一辆蓝色的轿车和一辆黑色的轿车。道路旁有低矮的灌木丛，为画面增添了一丝生活气息。道路的护栏和远处的灯光也构成了画面的层次感...

Hallusion

CapRL-3B

.....

*3. **Structure and Architecture:***
- The tower has a green roof with multiple tiers, typical of traditional Chinese architecture, specifically the pagoda style. It features a large, circular base with arches at the bottom, illuminated in gold, suggesting it is lit up at night.

*4. **Text on the Image:***
- At the top left of the image, there are large Chinese characters "长安," clearly indicating the location as Chang'an.

*5. **Textual Elements:***
*- The top right text reads **"遥观今时景 一梦回盛唐,"** which translates to "Looking at the scene today, a dream returns to the Tang Dynasty." This suggests a theme of revisiting or appreciating the historical significance of the place.*

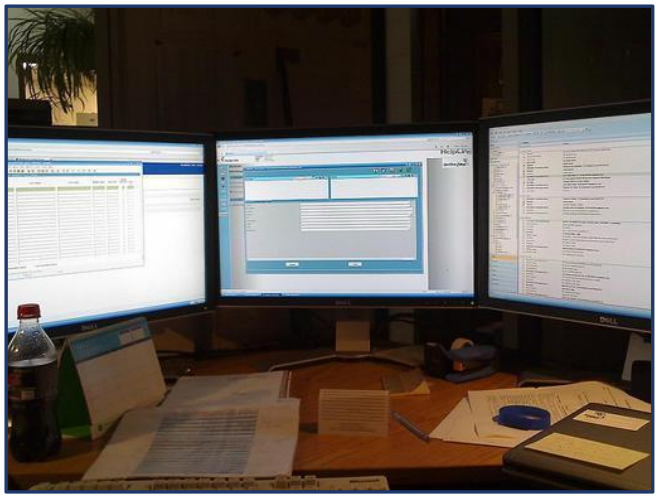
.....

Hallusion

Figure 5. A case for the caption OCR capability in CCCaption-2B, Qwen3-VL 2B, and CapRL-3B

nation detection due to its strict Yes/No question format.

- MMMU [23]: A massive multimodal understanding benchmark that spans college-level tasks across disciplines like Art, Science, and Engineering. Its complexity makes it a strong test for hallucination, as models must generate precise.
- MMStar [3]: An elite, vision-indispensable benchmark comprising six core capabilities. It uses a curated set of samples in which visual information is strictly necessary to answer the question.



Qwen3-VL-2B

This is a detailed description of the image:

The image captures a dimly lit office workspace, likely at night, with three computer monitors arranged in a triangle on a wooden desk. The monitors are all displaying the same software interface, which appears to be a data m ement or database application. The screens show a gri ata with columns and rows, suggesting the user analyzing information (Missing details).

*The monitors are Dell brand, as indicated by the logo on the botto . **The central monitor shows a form with a blue header and a white body, while the other two screens display similar data tables.** The workspace is cluttered with various office supplies, including a plastic soda bottle, **a blue cap**, a notepad, a pen, and a stapler. There is also a small, white, rectangular object that could be a small electronic device or a piece of paper.*

The room's lighting is minimal, with the monitors being the primary source of light. In the background, a potted plant is visible, adding a touch of greenery to the otherwise dark setting. The overall atmosphere is one of a dedicated, focused work environment.

Forgetfulness

Hallusion

CCCaption-2B

.....

*- **The Monitors:** There are three monitors arranged in a triangular formation on the desk. The monitor on the left displays a spreadsheet or a data table with numerous rows and columns. The middle monitor shows a window with a blue and white interface, possibly a software application or a form. The monitor on the right displays a large text-based interface, likely a log, a list, or a document with many lines of text. All three monitors are branded with the "DELL" logo.*

*- **The Desk and Surroundings:** The desk is made of wood and is cluttered with various items. In the foreground, there is a dark-colored soda bottle with a red cap, partially visible on the left. Next to it is a small, white and green calendar or notepad. The desk is also covered with papers, some of which are folded or partially obscured. A blue, circular object, possibly a lid or a small container, is visible on the right side of the desk. A black stapler and a pen are also present. A black keyboard is visible at the bottom of the frame.*

*- **Background and Lighting:** The background is dark, with a blurred view of a plant in the upper left corner and a doorway or window in the upper right. The lighting is low, with the monitors being the primary source of light, casting a bright glow on the desk and creating a contrast with the surrounding darkness. The overall atmosphere is one of a busy, focused work environment.....*

Qwen3-VL-32B

.....

*- **Papers and Documents:***

- Several printed documents are scattered across the desk.

- Some appear to be reports or forms with lines of text, possibly corresponding to the data on the screens.

- A small stack of paper with a hole punch is near the center.

*- **Office Supplies:***

*- A **blue tape dispenser** (likely for adhesive tape) is on the right side.*

*- A **pen** lies on the desk near the*

*- A **notebook or small notebook** w*

on top is on the far right.

*- **Beverage:***

- A **plastic bottle of soda** (possibly Coca-Cola, judging by the red cap and label design) is on the left

Nitpicking

Too much detail leads to caption truncation

Figure 6. A case for the complex scene summarization capability in CCCaption-2B, Qwen3-VL 2B, and Qwen3-VL 32B

CapArena [4] constructs an "arena-style" evaluation platform for granular assessment. It consists of 600 pairwise caption battles in which model outputs are pitted against each other, enabling a comparative analysis of caption quality, detail, and factual accuracy.

Prompts. We detail all MLLMs prompts employed within our framework in Figures 1 through 4. These include the prompts utilized for the captioning model, the completeness judge model, the correctness judge model, and the multi-MLLM production pipeline.

Experiments

Full Results. We provide the full evaluation results for prism and hallucination assessment in Tables 1 and 2. Overall, the CCCaption framework achieves state-of-the-art (SOTA) performance across multiple critical evaluation frameworks. On the prism evaluation, as shown in Table 1, CCCaption-2B improves upon CapRL 3B by 1.72 points (3.38%). Despite its smaller model size, CCCaption-2B performs comparably to the much larger Qwen3-VL 32B, maintaining a slight lead of 0.54 points. Notably, the

model achieves a significant improvement on OCRBench, scoring 55.34 and surpassing the second-place InternVL3.5 38B by 4.53 points (an 8.91% margin). Regarding hallucination assessment, as detailed in Table 2, CCCaption-2B demonstrates superior correctness, improving by 0.38 points (0.54%) over the base model Qwen3-VL 2B. In contrast, CapRL, which lacks an explicit correctness reward mechanism, shows a significant increase in hallucination compared to its base model, Qwen2.5-VL 3B, resulting in a decrease of 0.86 points (1.27%). Furthermore, CCCaption-2B exhibits higher correctness than Qwen3-VL 32B, with an improvement of 0.70 points (1.00%).

Case Analysis. We provide further examples illustrating the captions generated by our CCCaption-2B. As illustrated in Figure 5, a qualitative case study focusing on the caption OCR capability reveals that CCCaption-2B correctly transcribes and interprets the complex, stylized text within the image. In contrast, baseline models like Qwen3-VL 2B and CapRL-3B suffer from specific errors, such as excessive nitpicking and hallucination, respectively.

As illustrated in Figure 6, the assessment of complex scene summarization reveals clear distinctions among the models. CCCaption-2B provides a comprehensive, structured description of the cluttered office environment, whereas the baseline models, Qwen3-VL 2B and Qwen3-VL 32B, exhibit significant errors, including hallucinations, forgetting, and excessive nitpicking.

Further Discussion. To address potential concerns regarding evaluation bias and preference overfitting inherent in model-based judging, we validate our CCCaption framework using a diverse set of metrics beyond the Qwen3-VL-32B judge used in CapArena, including standard Hallucination Evaluation and Prism metrics, which confirm that performance gains stem from genuine improvements in completeness and correctness rather than reward hacking.

Regarding the computational overhead of the reinforcement learning pipeline, while calculating the correctness reward via MLLM-based verification incurs additional cost compared to standard SFT, our proposed Dynamic Query Sampling strategy significantly enhances training efficiency by prioritizing diverse and high-advantage queries, thereby mitigating the overall training burden. Furthermore, to ensure the robustness of the correctness validator and minimize the risk of reinforced hallucinations from false-positive rewards, we employ a decomposition strategy that simplifies the verification task into atomic sub-caption queries, making the process more tractable for the validator model; this is balanced by our dual-reward mechanism where the correctness penalty naturally counteracts the potential for excessive verbosity or repetition that might arise from the completeness objective, resulting in captions that are detailed yet factually grounded.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 4

[2] Fu Chaoyou, Chen Peixian, Shen Yunhang, Qin Yulei, Zhang Mengdan, Lin Xu, Yang Jinrui, Zheng Xiawu, Li Ke, Sun Xing, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 3, 2023. 4

[3] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *Advances in Neural Information Processing Systems*, 37:27056–27087, 2024. 4, 5

[4] Kanzhi Cheng, Wenpo Song, Jiaxin Fan, Zheng Ma, Qiushi Sun, Fangzhi Xu, Chenyang Yan, Nuo Chen, Jianbing Zhang, and Jiajun Chen. Caparena: Benchmarking and analyzing detailed image captioning in the llm era. *arXiv preprint arXiv:2503.12329*, 2025. 2, 6

[5] Mingqian Feng, Yunlong Tang, Zeliang Zhang, and Chenliang Xu. Do more details always introduce more hallucinations in lvlm-based image captioning? *arXiv preprint arXiv:2406.12663*, 2024. 4

[6] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14375–14385, 2024. 4

[7] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 1

[8] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 4

[9] Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. In *European conference on computer vision*, pages 235–251. Springer, 2016. 4

[10] Saehyung Lee, Seunghyun Yoon, Trung Bui, Jing Shi, and Sungroh Yoon. Toward robust hyper-detailed image captioning: A multiagent approach and dual evaluation metrics for factuality and coverage. *arXiv preprint arXiv:2412.15484*, 2024. 2, 4

[11] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Pro-*

ceedings of the IEEE/CVF conference on computer vision and pattern recognition, pages 26296–26306, 2024. 4

[12] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer, 2024. 4

[13] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024. 4

[14] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pages 2263–2279, 2022. 3, 4

[15] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022. 3, 4

[16] Thao Nguyen, Samir Yitzhak Gadre, Gabriel Ilharco, Sewoong Oh, and Ludwig Schmidt. Improving multimodal datasets with image captioning. *Advances in neural information processing systems*, 36:22047–22069, 2023. 4

[17] Runqi Qiao, Qiuna Tan, Peiqing Yang, Yanzi Wang, Xiaowan Wang, Enhui Wan, Sitong Zhou, Guanting Dong, Yuchen Zeng, Yida Xu, et al. We-math 2.0: A versatile mathbook system for incentivizing visual mathematical reasoning. *arXiv preprint arXiv:2508.10433*, 2025. 4

[18] Yuxuan Qiao, Haodong Duan, Xinyu Fang, Junming Yang, Lin Chen, Songyang Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. Prism: A framework for decoupling and assessing the capabilities of vlms. *Advances in Neural Information Processing Systems*, 37:111863–111898, 2024. 2, 3, 4

[19] Sheldon M Ross, Sheldon M Ross, Sheldon M Ross, Sheldon M Ross, and Etats-Unis Mathématicien. *A first course in probability*. Prentice Hall Upper Saddle River, NJ, 1998. 1

[20] Qwen Team. Qwen3 technical report, 2025. 4

[21] Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, et al. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Advances in Neural Information Processing Systems*, 37:113569–113697, 2024. 3, 4

[22] Long Xing, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Jianze Liang, Qidong Huang, Jiaqi Wang, Feng Wu, and Dahua Lin. Caprl: Stimulating dense image caption capabilities via reinforcement learning. *arXiv preprint arXiv:2509.22647*, 2025. 3, 4

[23] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9556–9567, 2024. 4, 5

[24] Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15134–15186, 2025. 4

[25] Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Yu Qiao, et al. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? In *European Conference on Computer Vision*, pages 169–186. Springer, 2024. 3, 4