DQ-RL: FINE-GRAINED REWARDS FROM DIVERSE QUERIES FOR DENSE IMAGE CAPTIONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Image captioning is a fundamental task in vision-language understanding and a core capability of large vision-language models. Existing efforts to improve this capability mainly rely on supervised fine-tuning of base models, often on domain-specific data, while recent work has begun to explore reinforcement learning (RL) for better generalization. However, applying RL to image captioning remains challenging: as an open-ended generation task, captioning lacks verifiable rewards, making reward design under the Reinforcement Learning with Verifiable Rewards (RLVR) paradigm particularly difficult. To address this challenge, we propose **Diverse-Query Reinforcement Learning** (**DQ-RL**). We design completeness queries (**CQs**) and faithfulness queries (**FQs**), perform unified sampling over the candidate query pool, evaluate sub-capabilities of generated captions, and use the resulting quantitative scores as rewards for optimization. Following this paradigm, we construct a 40k training dataset named **DQ-CAP-40k**. Extensive experiments show that DQ-RL, a 4B model, achieves an average improvement of 1.4% over baseline models under the PRISM evaluation framework, and matches Qwen3.5-27B model at 63.1%. We also build **DQ-Bench**, an evaluation benchmark tailored for image captioning. On the three core metrics of DQ-Bench, namely Correct, Omit and Wrong, DQ-RL outperforms the baseline by 8.6%, 7.3%, and 1.3%, respectively.

1 INTRODUCTION

Image captioning (Vinyals et al., 2015; Karpathy & Fei-Fei, 2015) is a fundamental vision-language task, which requires models to generate natural language descriptions that accurately characterize the visual content of input images (Lu et al., 2019; Radford et al., 2021). Currently, large vision-language models (LVLMs) are widely adopted to tackle the image captioning task. Representative LVLMs including GPT (Singh et al., 2025), Gemini (Team et al., 2023), LLaVA (Liu et al., 2023), and QwenVL (Bai et al., 2023) have achieved substantial improvements in image-text alignment after training on web image-caption pairs (Schuhmann et al., 2022; Kirillov et al., 2023). Although powerful LVLMs such as Gemini can generate high-quality captions, they are often costly and inefficient for practical deployment. In contrast, small-scale open-source LVLMs are more efficient but still lag behind in captioning quality.

In general, the quality of generated image captions is positively correlated with the capability and model scale of LVLMs, as our evaluations confirm in Table 2. For practical deployment, open-source models (Team, 2026) offer prominent advantages in inference cost and efficiency and are widely adopted for relevant tasks. Nevertheless, small-scale variants (2B/4B/9B) exhibit substantially weaker performance compared with larger models. Therefore, boosting the captioning capabilities of small-scale LVLMs is of great necessity. To enhance such capabilities, prior works (Chen et al., 2024; Rotstein et al., 2024; Vasu et al., 2025) adopt supervised fine-tuning (SFT), where model performance is improved by elevating training data quality. However, these approaches suffer from limited generalization due to constraints on data scale. Recently, reinforced by the success of RLVR in LLM research, several studies have explored applying RLVR to the image captioning task (Tang et al., 2026; Xing et al., 2026). Conventional RLVR for question answering (QA) only assigns reward scores based on the final answer. In contrast, image captioning is an open-ended task that requires evaluating the entire generated output. Wang et al. (2025) adopt the LLM-as-judge paradigm to assign a single reward score for the full caption. However, as LVLMs become more ca-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

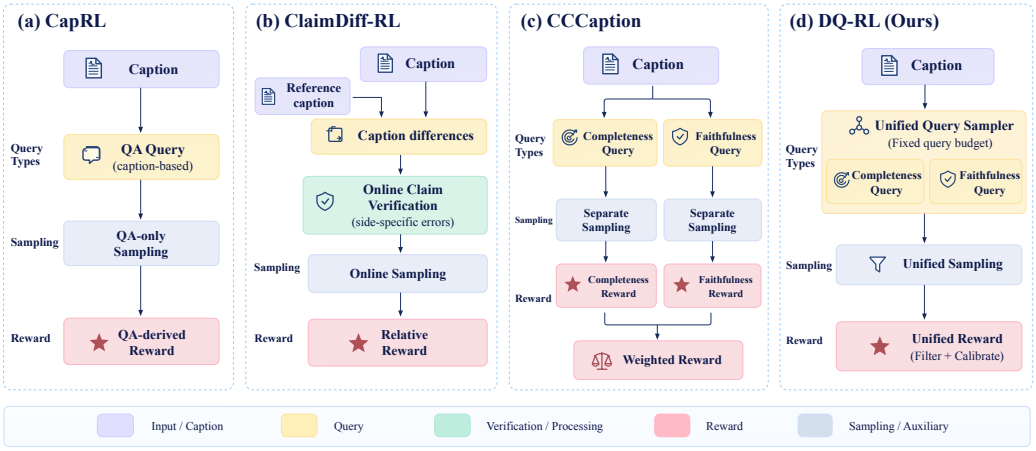


Figure 1: Comparison of query-based RL methods. (a) **CapRL** uses only QA queries; (b) **ClaimDiff-RL** performs online claim verification for side-specific errors; (c) **CCCaption** separately handles completeness and faithfulness queries; (d) **DQ-RL** jointly samples both queries with a unified sampling and reward strategy.

pable, holistic scoring tends to yield inflated scores with poor discriminability, which renders RLVR algorithms such as GRPO (Shao et al., 2024) ineffective.

To address the drawbacks of holistic scoring, Xing et al. (2026) performs RLVR using multiple-choice QA accuracy for image captions. It reduces omission and improves comprehensiveness, but struggles to penalize erroneous content, increasing hallucination. Li et al. (2026) instead proposes fine-grained RL for captioning, using reference-conditioned atomic visual claim differences as rewards. However, its online decomposition and full evaluation rely on closed-source LVLMs as judges, making training prohibitively expensive. These works move caption evaluation from holistic scoring toward decomposed fine-grained assessment. Building on this direction while addressing their limitations, we treat the image–caption pair as evidence and design queries for individual capability sub-items. We focus on two indispensable dimensions: completeness and faithfulness—a high-quality caption should cover all key objects and scenes while remaining precise and hallucination-free. Each query yields an independent fine-grained evaluation signal, which is further utilized as the supervision reward for reinforcement learning. Since completeness queries and faithfulness queries provide two distinct reward signals, we introduce a unified sampling strategy to jointly sample them during RL training. Accordingly, we term our method **Diverse-Query Reinforcement Learning (DQ-RL)**. As shown in Figure 1, we clearly demonstrate the differences among CapRL, ClaimDiff-RL, CCCaption and DQ-RL.

We now detail the construction of the two query types in DQ-RL. We first focus on completeness. For image captioning, completeness corresponds to recall: the caption should cover sufficient visual content without omitting critical information. However, completeness is inherently subjective. Prior works (Liu et al., 2025; Tong et al., 2025) therefore predefine an evaluation system and verify each item individually. Automated caption evaluation relies on powerful LVLMs for reliable judgments. As shown in Table 5, stable and credible evaluation heavily depends on high-capacity LVLMs, which incur substantially higher computational costs. Since RL training involves iterative rollouts, directly using powerful judge models leads to prohibitive overhead. To reduce this cost, we follow Xing et al. (2026) and construct offline query sets as completeness queries (CQs). We further design these queries to target and compensate for the completeness deficiencies in the base model's outputs.

We then turn to faithfulness queries (FQs). Beyond completeness, faithfulness requires generated captions to be factually correct and hallucination-free, which cannot be assessed offline—a key limitation of Xing et al. (2026). Tang et al. (2026) addresses this by decomposing atomic capabilities via sub-captions, but when the base model is strong, most sub-captions are correct, yielding weak supervision. Inspired by this idea, we introduce semantic statements, which lie between holistic scoring and atomic capability evaluation. They aggregate multiple atomic capabilities while avoiding complicated holistic scoring, and their scores only account for erroneous content, alleviating the low discriminability of optimization signals for strong base models. We decompose each rollout

caption into FQs at semantic-statement granularity. We construct **DQ-CAP-40k** from 40k images, each paired with multiple offline CQs and their reference answers. During RL training, FQs are generated online from each rollout caption.

For evaluation, we adopt PRISM (Qiao et al., 2024) and construct **DQ-Bench**, containing 2k images across 4 major categories and 20 subcategories. We build a fully LVLM-based evaluation pipeline to guide model iteration and selection.

In summary, our contributions are three-fold:

- We propose DQ-RL, a fine-grained reinforcement learning framework that decomposes image caption evaluation into completeness queries and faithfulness queries, improving caption completeness and faithfulness, respectively. We further introduce a unified sampling strategy to jointly optimize the model with multiple reward signals.
- We construct DQ-CAP-40k for RL training and DQ-Bench for image captioning evaluation. DQ-CAP-40k contains 40k samples, while DQ-Bench covers 2k images across 4 major categories and 20 subcategories.
- Extensive experiments demonstrate the effectiveness of DQ-RL. It outperforms baseline methods under both PRISM and DQ-Bench, and even surpasses larger models, showing the potential of query-based RL for small-scale LVLMs.

2 Related Works

Image Captioning. Image captioning has evolved from early encoder-decoder and attention-based models (Vinyals et al., 2015; Karpathy & Fei-Fei, 2015; Xu et al., 2015) to large vision-language models such as BLIP, BLIP-2, and LLaVA (Li et al., 2022; 2023; Liu et al., 2023). Recent surveys provide comprehensive overviews of this progress (Singh et al., 2024; Sharma et al., 2026). Beyond generic captioning, recent work has extended the task toward controllable language patterns (Saito et al., 2025) and detailed localized descriptions (Lian et al., 2025). In contrast to region-level dense captioning (Johnson et al., 2016), which generates multiple local captions, detailed whole-image captioning aims to produce a single caption that covers a broad range of visual facts (Hua et al., 2025; Lin et al., 2025; Hirota et al., 2025; Lu et al., 2025). However, how to provide fine-grained and reliable supervision for such detailed captions remains underexplored.

Reinforcement Learning for Captioning. Self-critical sequence training established sequence-level reinforcement learning for image captioning by directly optimizing caption metrics such as CIDEr (Rennie et al., 2017). Recent RL-based captioning methods construct more explicitly verifiable reward signals. CapRL (Xing et al., 2026) uses image-grounded questions answered from a generated caption to measure information recovery. SC-Captioner (Zhang et al., 2025) constructs decomposed feedback over objects, attributes, and relations for iterative self-correction. CCCaption (Tang et al., 2026) combines complementary rewards for caption completeness and correctness. ClaimDiff-RL (Li et al., 2026) compares an actor caption with a reference caption and verifies their fine-grained claim differences against the image. In contrast, DQ-RL coordinates completeness and faithfulness at the query level: it uses fixed offline CQs for completeness and FQs for faithfulness, samples them via a unified strategy, and aggregates their scores in a single RL objective.

Evaluation for Image Captioning. Image captioning has traditionally been evaluated with reference-based metrics such as BLEU, METEOR, CIDEr, and SPICE (Papineni et al., 2002; Banerjee & Lavie, 2005; Vedantam et al., 2015; Anderson et al., 2016). Embedding- and model-based metrics, including CLIPScore and CLAIR (Hessel et al., 2021; Chan et al., 2023), extend evaluation beyond surface-form overlap by assessing semantic compatibility between images and captions. Fine-grained metrics and benchmarks further examine specific aspects of visual content. CHAIR measures object hallucination (Rohrbach et al., 2018); CAPability evaluates caption correctness and thoroughness across multiple visual dimensions (Liu et al., 2025); and CaptionQA measures whether a caption preserves image-level utility through image-grounded questions (Yang et al., 2026). Building on this diagnostic perspective, we construct DQ-Bench, a fine-grained evaluation set that decomposes caption quality into completeness and faithfulness, measuring correct descriptions, omissions, and incorrect statements. Here, omissions reflect completeness, while incorrect statements reflect faithfulness/hallucination. However, existing benchmarks mainly evaluate final captions and are not designed to provide query-level reward signals for RL training.

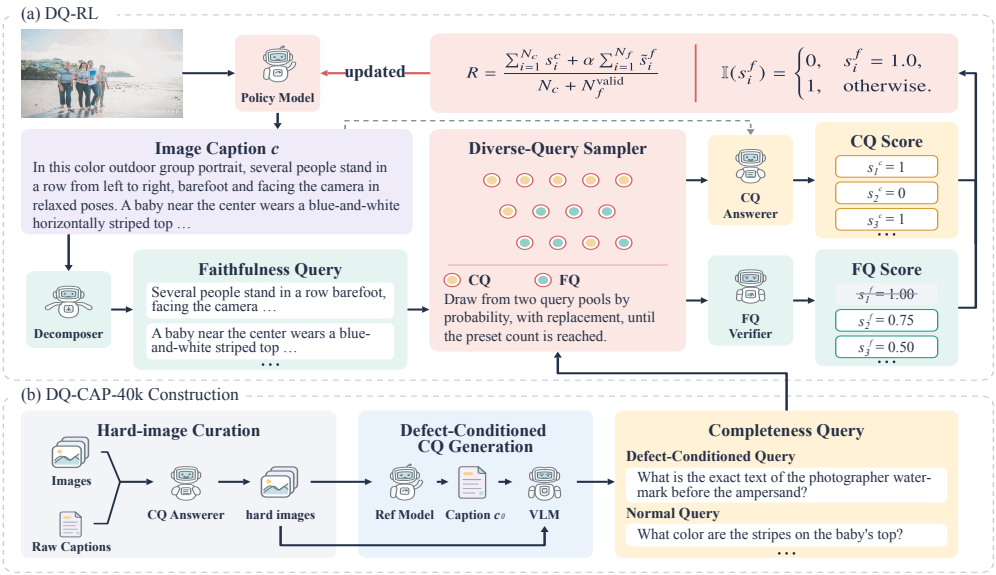


Figure 2: Overview of DQ-RL.(a) Offline completeness queries (CQs) and online faithfulness queries (FQs) jointly evaluate each rollout. A unified sampler draws from both query pools under a fixed budget, and the resulting scores are aggregated into a single reward.(b) DQ-CAP-40k augments general CQs with hard CQs targeting omissions and errors in base-model captions.

3 METHODS

Figure 2(a) provides an overview of the reinforcement learning pipeline of **Diverse-Query Reinforcement Learning (DQ-RL)**. Given an image, the VLM generates a textual description of its visual content. To optimize caption quality, we compute rewards using two types of queries: faithfulness queries, which encourage the caption to be factually faithful to the image and thereby improve precision, and completeness queries, which encourage the caption to cover comprehensive visual content and thereby improve recall. In this section, we first introduce the design of faithfulness queries (Section 3.1.2) and completeness queries (Section 3.1.1), and then detail the reward functions (Section 3.3).

3.1 DIVERSE QUERY

Image Caption. A VLM policy π_θ defines a conditional distribution over caption sequences. Given an input image I and a prompt $\boldsymbol{q}$, the policy generates a caption $c = (c^1, \ldots, c^T)$ token by token:

$$\pi_\theta(c \mid I, \boldsymbol{q}) = \prod_{t=1}^{T} \pi_\theta(c^t \mid c^{<t}, I, \boldsymbol{q}), \tag{1}$$

where c^t denotes the t-th generated token and $c^{<t} = (c^1, \ldots, c^{t-1})$ denotes the preceding token prefix. In DQ-RL, we optimize π_θ with diverse query supervision to improve the faithfulness and completeness of the generated caption c.

Diverse Queries. Diverse queries are key to precise caption evaluation. Directly assigning an overall score to a lengthy descriptive caption is difficult and subjective. Inspired by human evaluation practices, we build an evaluation system with predefined dimensions and criteria. Such a system decomposes caption quality into atomic dimensions. Although atomic dimensions enable precise evaluation, they suffer from over-decomposition and sharply increase the number of evaluation items. To balance precision and efficiency, we construct image-grounded question–answer pairs offline, each serving as a completeness query (CQ). For each rollout, we decompose its caption into faithfulness queries (FQs) at semantic-statement granularity. Each FQ groups related visual details into a verifiable unit between atomic dimensions and the full caption. Together, CQs and FQs provide complementary supervision for caption completeness and faithfulness.

3.1.1 COMPLETENESS QUERIES

As discussed above, completeness evaluation requires judging whether a caption covers sufficient visual facts. Automated evaluation under such a system is typically implemented with powerful LVLMs. Evaluation reliability generally improves with model capacity, and closed-source LVLMs often provide more reliable judgments than open-source alternatives. For example, ClaimDiff-RL (Li et al., 2026) adopts Gemini as the judge model. However, standard RL algorithms such as GRPO require repeated rollouts followed by evaluation, making it prohibitively expensive to invoke such strong models at every training step. To avoid this overhead, we construct completeness queries offline and reuse them throughout training.

Specifically, for hard-query curation, we employ a powerful LVLM G with a carefully designed prompt P_G. Given the image I and a caption c_0 generated by the base model, G produces additional hard queries and their reference answers:

$$\mathcal{Q}^{\text{hard}}(I) = \{(q_i^h, a_i^h)\}_{i=1}^{n_h} = G(P_G, I, c_0). \tag{2}$$

Each hard query targets a visual fact omitted or misstated in c_0, so that these queries address deficiencies in the base-model caption. The hard query–answer pairs are combined with the general CQs to form the offline completeness-query pool $\mathcal{Q}^{\text{comp}}(I)$, which is fixed during RL training and used to compute completeness rewards without repeatedly invoking G.

3.1.2 FAITHFULNESS QUERIES

In contrast to completeness, faithfulness targets factual correctness: the generated caption should align precisely with the image content and avoid hallucinations or factual errors. We decompose each rollout caption into FQs at semantic-statement granularity and verify each FQ against the image individually. This makes long captions easier to evaluate while retaining related visual details within each query.

Given an image I and a caption c generated by the policy, a text-only LLM G_S with prompt P_S produces the FQs:

$$[f_1, f_2, \ldots, f_{n_f}] = G_S(c, P_S), \tag{3}$$

where f_i denotes the i-th FQ and n_f is the number of FQs generated for the current rollout. Unlike CQs, which are constructed offline and fixed during training, FQs depend on the current rollout caption c and are therefore generated online. Each FQ is then verified against the image I to produce a faithfulness score used in the reward.

3.2 UNIFIED QUERY SAMPLING

For an image I, the completeness queries are constructed offline and fixed during training. Given a rollout caption c, we further decompose it into faithfulness queries online. We combine both types of queries into a single pool and perform unified query sampling. Specifically, the candidate pool contains n_c CQs and n_f FQs:

$$\mathcal{Q}(I, c) = \mathcal{Q}^{\text{comp}}(I) \cup \mathcal{Q}^{\text{faith}}(c), \tag{4}$$

where $\mathcal{Q}^{\text{comp}}(I) = \{(q_i, a_i)\}_{i=1}^{n_c}$ is the offline CQ pool and $\mathcal{Q}^{\text{faith}}(c) = \{f_i\}_{i=1}^{n_f}$ is the online FQ pool. We then sample N_q queries from this candidate pool.

We implement the heterogeneous query distribution as a two-stage sampler. A routing variable first selects the query source, after which a query is drawn from the corresponding collection:

$$z \sim \text{Bernoulli}(\lambda), \qquad x \sim \begin{cases} p(\cdot \mid \mathcal{Q}^{\text{comp}}(I)), & z = 1, \\ p\big(\cdot \mid \mathcal{Q}^{\text{faith}}(c)\big), & z = 0. \end{cases} \tag{5}$$

Here λ is the probability of selecting a completeness query. Given z, the query x is drawn from the corresponding query set: if $z = 1$, one completeness query is selected from $\mathcal{Q}^{\text{comp}}(I)$; if $z = 0$, one faithfulness query is selected from $\mathcal{Q}^{\text{faith}}(c)$. Thus, λ controls the allocation between CQs and FQs. Among the N_q sampled queries, N_c are CQs and N_f are FQs, so $N_c + N_f = N_q$.

3.3 REWARD SYSTEM DESIGN

With the constructed diverse queries, we build a unified VLM/LLM-as-Judge evaluation system to assess caption quality. Given an image I and its caption c, let $Q = Q^{\text{comp}} \cup Q^{\text{faith}}$ denote the set of sampled queries, where Q^{comp} and Q^{faith} are completeness and faithfulness queries, respectively.

For each query $q_i \in Q$, a judge model J produces a score s_i that measures the consistency between the image I and caption c:

$$s_i = J(I, c, q_i). \tag{6}$$

For each sampled query $q_i \in Q = Q^{\text{comp}} \cup Q^{\text{faith}}$, the judge model J produces a score $s_i = J(I, c, q_i) \in [0, 1]$ following Eq. equation 6. For a completeness query $q_i \in Q^{\text{comp}}$, s_i reflects whether the visual information targeted by the query is correctly covered by the caption c. For a faithfulness query $q_i \in Q^{\text{faith}}$, s_i reflects whether the FQ is supported by the image I. Since both types of queries share the same judge model and produce scores in the same range, their scores can be naturally aggregated into a unified reward for reinforcement learning.

In current multimodal large models, image captioning capability is already fairly strong (see Table 2, where the base model achieves high metrics), and hallucination or factual errors are relatively rare. As a result, many faithfulness queries receive a full score, which pushes up the average faithfulness reward and weakens its discriminability. To address this issue, we filter out faithfulness queries that are already completely correct and introduce an indicator function

$$\mathbb{I}(s_i^f) = \begin{cases} 0, & s_i^f = 1.0, \\ 1, & \text{otherwise.} \end{cases} \tag{7}$$

The effective score of a faithfulness query is then computed as $\tilde{s}_i^f = s_i^f \times \mathbb{I}(s_i^f)$, and the number of valid faithfulness queries is $N_f^{\text{valid}} = \sum_{i=1}^{N_f} \mathbb{I}(s_i^f)$. For a caption obtained from one rollout, after dynamically sampling queries and computing the reward score for each query, the reward is

$$R = \frac{\sum_{i=1}^{N_c} s_i^c + \alpha \sum_{i=1}^{N_f} \tilde{s}_i^f}{N_c + N_f^{\text{valid}}}, \tag{8}$$

where s_i^c and s_i^f denote the scores of completeness and faithfulness queries, respectively, and α is a balancing coefficient that controls the relative importance of the two types of queries. When $N_f^{\text{valid}} = 0$, the faithfulness term is set to zero.

3.4 Data Construction

3.4.1 Training Dataset

We collect and filter 40k images for training. Following the label schema in Wu et al. (2025), the images cover four domains: Nature, Design, People, and Synthetic. Figure 7 shows the detailed distribution. The overall distribution is balanced, which helps improve the model's general capability.

After collecting valid images, we construct completeness queries for each image. Our construction follows the QA Curation method in Xing et al. (2026). For each image, a VLM with carefully designed prompts generates queries and answers. We focus on perceptual capability, so all queries are required to be directly observable from the image.

Hard CQ Curation. The above pipeline is general and effective, but many samples are relatively simple, leading to high initial completeness rewards (exceeding 0.6). To address this, we introduce hard completeness queries (hard CQs), as illustrated in Fig. 2(b). Given a base model π, we first generate a caption c_0 for each training image. A powerful multimodal model then uses the image I and caption c_0 to produce targeted queries that focus on missing details and hallucinated content in c_0, increasing query quality and difficulty. Finally, multiple VLMs verify each query to ensure correctness.

3.4.2 Test Benchmark

We collect a 2k-image evaluation set. Following the same label schema, we uniformly sample images within each subcategory and filter out those with abnormal resolution, blurriness, or other artifacts. The resulting benchmark maintains satisfactory quality and balanced distribution. We further design a fine-grained diagnostic protocol to examine different aspects of caption quality.

Fine-grained Diagnostic Evaluation. We evaluate caption quality across 14 atomic fields in six categories: subject, quantity, color, composition, background, and visual style. For each image–caption pair, the evaluator compares the caption with the visual evidence and labels every applicable field as *Correct*, *Omit*, or *Wrong*. *Correct* means the visual information is accurately described; *Omit* means relevant information is present in the image but absent from the caption; *Wrong* means

the caption contains information inconsistent with or unsupported by the image. Fields absent from both the image and the caption are excluded.

Let $\Omega_{\mathcal{D}}$ denote all applicable evaluation records in dataset $\mathcal{D}$. For $k \in \{\text{correct}, \text{omit}, \text{wrong}\}$, we compute

$$\text{Rate}(k) = \frac{1}{|\Omega_{\mathcal{D}}|} \sum_{r \in \Omega_{\mathcal{D}}} \mathbb{I}[y_r = k]. \tag{9}$$

The three labels are mutually exclusive, so their rates sum to one. We report both overall results and category-level breakdowns.

4 EXPERIMENTS

We evaluate captions using two complementary, reward-independent protocols: DQ-Bench and PRISM (Qiao et al., 2024). DQ-Bench provides fine-grained diagnostics on completeness and faithfulness, while PRISM tests whether captions preserve information for downstream multimodal QA.

4.1 EVALUATION FRAMEWORK

DQ-Bench. We evaluate captions using two complementary, reward-independent protocols: our diagnostic benchmark measures visual coverage and factual support at the field level, while PRISM measures whether a generated caption preserves sufficient information for VQA.

Our diagnostic evaluator assigns each applicable evaluation field to one of three outcomes: *Correct*, *Omit*, or *Wrong*. *Correct* means the caption accurately describes the image content; *Omit* means relevant image content is absent from the caption; and *Wrong* means the caption contains unsupported or incorrect content. We report the three rates jointly. Since the three outcomes are mutually exclusive and cover all applicable fields, their rates sum to 100%. Correct reflects successful coverage, whereas Omit and Wrong correspond to the failure modes targeted by completeness and faithfulness queries, respectively. The evaluation fields, judgment rubric, and evaluator configuration are detailed in Sec. 3.4.2.

PRISM Evaluation. We adopt the released CapRL PRISM pipeline (Xing et al., 2026; Qiao et al., 2024) as a functional test of caption quality: a fixed Qwen2.5-VL-3B answerer, fine-tuned for caption-based QA, receives the generated caption and the benchmark question but never the image. We cover eight benchmarks: ChartQA (Masry et al., 2022), CharXiv (Wang et al., 2024), HallusionBench (Guan et al., 2024), MathVerse (Zhang et al., 2024), WeMath2.0-Pro (Qiao et al., 2026), MMBench (Liu et al., 2024a), MME (Fu et al., 2025), and OCRBench (Liu et al., 2024b).

4.2 IMPLEMENTATION DETAIL

Training configuration. All RL variants start from the same `Qwen3.5-4B` base policy and are trained on the same 40k images described in Sec. 3.4. In each ablation, we vary only the specified query-mixture factor and keep the remaining training settings fixed.

We fully fine-tune the captioning policy with verl and use Ray for distributed execution. Training uses FSDP2, bf16 precision, and gradient checkpointing. We generate caption rollouts asynchronously with vLLM at a tensor parallel size of 4. The optimizer is AdamW with an actor learning rate of 1×10^{-5}, a 20-step warmup followed by a constant learning rate, and a gradient clipping threshold of 1.0. Each training batch contains 128 images, with four rollout captions generated per image at a sampling temperature of 1.0. We use GRPO for policy optimization, set the KL coefficient to 0, and train for 3 epochs.

Reward configuration. Rewards are computed by a separate service. For each rollout caption, the unified sampler draws $N_q = 13$ queries from completeness queries and faithfulness queries. Each draw selects a completeness query with probability $\lambda = 0.5$ and a faithfulness query otherwise. Faithfulness-query scores are scaled by $\alpha = 0.75$.

4.3 MAIN RESULTS

Results on PRISM. Table 1 compares the evaluated captioners on the eight public PRISM benchmarks. Each model is used only as a captioner: a fixed second-stage answerer receives the generated caption and the benchmark question, but never the image. DQ-RL achieves the best performance among open-source models under the PRISM evaluation protocol, matching Qwen3.5-27B at 63.1%.

Table 1: PRISM caption-mediated evaluation results. Downstream QA accuracy (%) on eight benchmarks and average (higher is better). Bold/underline: best/second-best among open-source models per column. *: from CCCaption (Tang et al., 2026); all others reproduced by us under a unified pipeline. CapRL-4B uses Qwen3.5 with configurations identical to CapRL-3B. Model citations: Gemini (Team et al., 2023), Seed (Seed, 2026), Qwen (Team, 2026),CapRL Xing et al. (2026).

MLLM	ChartQA	CharXiv	Hallusion	Verse	MMB	MME	OCR	WM2Pro	Avg.
Gemini-3.0-flash	85.2	45.6	71.9	39.7	84.7	83.2	64.3	40.8	64.4
Seed-2.0-lite	87.6	46.3	70.0	36.5	83.0	82.7	59.9	41.4	63.4
Qwen3.5-27B	86.1	<u>47.5</u>	69.1	**39.8**	<u>83.0</u>	75.2	62.8	**41.0**	<u>63.1</u>
CCCaption*	75.1	37.0	–	35.1	79.3	–	55.3	39.8	–
CapRL-3B	85.2	40.3	67.7	31.5	81.7	**76.1**	59.8	37.9	60.0
CapRL-4B	85.5	43.2	69.5	35.1	**83.7**	<u>75.4</u>	<u>64.0</u>	39.2	62.0
Qwen3.5-4B	85.7	45.9	68.8	37.5	81.2	74.0	60.5	40.0	61.7
DQ-RL-4B	**87.2**	**48.3**	**69.7**	<u>38.6</u>	82.7	74.4	63.6	<u>40.4</u>	**63.1**

Table 2: Results of the same captioners on DQ-Bench. Correct, Omit, and Wrong are mutually exclusive rates (%; Correct higher is better; Omit and Wrong lower are better).

Metric	Gemini 3.0-Flash	Seed 2.0-lite	Qwen3.5 27B	CapRL 3B	CapRL 4B	Qwen3.5 4B	**DQ-RL-4B (Ours)**
Correct ↑	96.5	96.4	<u>91.8</u>	77.6	88.6	85.5	**94.1**
Omit ↓	2.8	2.6	5.4	9.5	<u>3.4</u>	9.4	**2.1**
Wrong ↓	0.7	1.0	**2.8**	12.9	7.9	5.1	<u>3.8</u>

Specifically, it obtains the best results on ChartQA and CharXiv, and the second-best performance on MathVerse, MMBench, and OCRBench. These results demonstrate that captions generated by DQ-RL are sufficiently informative for downstream VQA benchmarks, reflecting its strong perceptual capability.

Results on DQ-Bench. We then evaluate the same captioners on our diagnostic benchmark, which separates successful grounded recovery from omission and unsupported content. DQ-Bench evaluates caption quality over 14 atomic fields grouped into six categories, assigning each applicable field a *Correct*, *Omit*, or *Wrong* label. A detailed case-level example of this evaluation procedure is provided in Appendix A.6. As shown in Table 2, DQ-Bench decomposes generated captions into atomic visual fields and computes metrics after item-wise assessment. DQ-RL achieves substantial improvements over the baseline Qwen3.5-4B: Correct increases by 8.6%, Omit decreases by 7.3%, and Wrong decreases by 1.3%. Our method is also competitive against larger models. Compared with Qwen3.5-27B, our approach improves Correct by 2.3% and reduces Omit by 3.3%.

4.4 Ablation Study

We conduct ablation experiments on key hyperparameters and core designs to analyze the behavior of DQ-RL. The results validate its effectiveness and robustness.

Controlling the Query Mixture. For each rollout, we combine the image's offline CQs with FQs from the current caption to form a candidate query pool. During sampling, the probability coefficient λ controls the query mixture, and the importance factor α adjusts the unified reward. At a fixed $\alpha = 0.5$, routing more slots to CQs reduces Omit ($3.90\% \rightarrow 1.70\%$ from $\lambda = 0.3$ to 0.8) and raises Correct ($93.30\% \rightarrow 94.20\%$), while allocating fewer slots to faithfulness verification raises Wrong ($2.80\% \rightarrow 4.10\%$). The two single-source endpoints make this complementarity explicit: FQ-only training is conservative but omits information, whereas CQ-only training recovers content but admits more unsupported claims. Thus, λ selects a transparent completeness–faithfulness operating point by allocating a fixed verification budget between the two error types.

At equal routing, calibration selects the strongest balanced point. Increasing α from 0.25 to 0.75 raises Correct from 92.1% to 94.1% and reduces Omit from 2.5% to 2.1%, while retaining a 3.8% Wrong rate. In contrast, the unscaled online score ($\alpha = 1.0$) sharply increases Omit to 5.60%. Thus, routing and calibration play distinct roles: routing determines which error type receives verifier effort, whereas calibration determines how strongly unresolved FQs shape the final reward.

Table 3: Effects of query routing and score calibration on DQ-Bench. The first five columns vary the completeness-query routing probability λ (fixed $\alpha = 0.5$); the last four vary the faithfulness-score calibration factor α (fixed $\lambda = 0.5$). $\lambda = 0$ and $\lambda = 1$ denote FQ-only and CQ-only training.

	Routing ($\alpha = 0.5$)					**Calibration** ($\lambda = 0.5$)			
Metric	$\lambda = 0$	0.3	0.5	0.8	1.0	$\alpha = 0.25$	0.50	0.75	1.00
Correct ↑	89.4	93.3	93.8	**94.2**	92.2	92.1	93.8	**94.1**	91.6
Omit ↓	8.6	3.9	2.7	**1.7**	1.9	2.5	2.7	**2.1**	5.6
Wrong ↓	**1.9**	2.8	3.6	4.1	5.9	5.4	3.6	3.8	**2.8**

Table 4: Effect of full-score exclusion. NE: without exclusion.

	DQ-RL	NE
C ↑	**94.1**	92.6
O ↓	**2.1**	2.8
W ↓	**3.8**	4.6

Table 5: Evaluator sensitivity. G/S/Q: Gemini-3.1-Pro/Seed-2.0-Lite/Qwen3.5-9B.

	G	S	Q
C ↑	85.6	87.4	96.4
O ↓	9.4	8.8	1.1
W ↓	5.1	3.9	2.5

Table 6: Effect of the total query budget on answer outcomes. C/O/W: correct/omit/wrong.

Budget	2	4	8	**13**	16
C ↑	91.0	91.4	92.8	**94.1**	92.8
O ↓	5.7	3.5	4.10	**2.1**	5.0
W ↓	3.3	5.1	3.0	3.8	**2.3**

Effect of Full-Score Query Exclusion. As described in Sec. 3.3, we exclude faithfulness queries with a score of 1.0 when designing the reward system. We conduct an ablation study to examine this design choice. As shown in Table 4, using all faithfulness queries for reward computation negatively affects Correct, Omit, and Wrong: Correct drops from 94.1% to 92.6%, Omit rises from 2.1% to 2.8%, and Wrong increases from 3.8% to 4.6%. We attribute this to the strong capability of the base model, which tends to produce high scores on faithfulness queries. Retaining all such queries weakens the discriminative signal for optimizing captions.

Sensitivity to the Diagnostic Evaluator. The Correct/Omit/Wrong protocol decomposes caption quality into interpretable failure modes, but their absolute rates depend on evaluator calibration. We quantify this sensitivity on captions generated by the untrained Qwen3.5-4B policy. Holding the evaluation units, target evidence, and judgment protocol fixed, Table 5 compares different evaluators. The judges induce different absolute operating points: Correct ranges from 85.6% to 96.4%. This check isolates evaluator calibration rather than captioner performance; accordingly, all headline diagnostic comparisons use one fixed evaluator and one fixed protocol for every model.

Effect of Query Budget. Table 6 studies the effect of the total query budget. Increasing the budget improves performance up to $N_q = 13$, which achieves the highest Correct and lowest Omit rates. Increasing the budget to 16 yields no consistent gains, suggesting saturation beyond a moderate budget. We therefore use $N_q = 13$ in the main experiments.

4.5 Qualitative Analysis

We present qualitative comparisons among Qwen3.5-4B (the base model), CapRL-4B, Qwen3.5-27B, and DQ-RL in Figure 4. Captions generated by Qwen3.5-4B evidently omit many fine-grained details, while CapRL-4B suffers from substantially increased hallucinations. In contrast, DQ-RL achieves superior performance in both completeness and faithfulness. Additional qualitative examples are provided in Figures 5 and 6.

5 Conclusions

In this work, we propose Diverse-Query Reinforcement Learning (DQ-RL) to jointly improve completeness and faithfulness in detailed image captioning. DQ-RL combines completeness queries, which target visual information missed by the base policy, with faithfulness queries, which verify claims in generated captions. Unified query sampling and discriminative reward aggregation provide complementary supervision for recovering image content and correcting unsupported descriptions. We further construct DQ-CAP-40k for training and DQ-Bench for fine-grained evaluation. Experiments show that DQ-RL reduces both omissions and factual errors relative to the base policy and improves caption-mediated question answering. Notably, our 4B model performs comparably to Qwen3.5-27B under the PRISM evaluation, despite being much smaller. These results demonstrate the effectiveness of query-based rewards for improving the captioning capabilities of smaller vision-language models.

AI USAGE DISCLOSURE

In this work, generative AI tools were used to improve the English language, grammar, and readability of the manuscript. They were also used to assist with code development during the implementation of the proposed method. All AI-assisted code was carefully inspected, modified where necessary, and tested by the authors to verify its correctness. The authors take full responsibility for the final manuscript, code, experimental results, and conclusions.

REFERENCES

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. Spice: Semantic propositional image caption evaluation. In *European conference on computer vision*, pp. 382–398. Springer, 2016.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023. URL https://arxiv.org/abs/2308.12966.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72, 2005.

David M Chan, Suzanne Petryk, Joseph E Gonzalez, Trevor Darrell, and John Canny. Clair: Evaluating image captions with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13638–13646, 2023.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, pp. 370–387. Springer, 2024.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. Mme: A comprehensive evaluation benchmark for multimodal large language models. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025. doi: 10.52202/085713-4899. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/d79a27cf2772fe00be7f341efc0eb517-Paper-Datasets_and_Benchmarks_Track.pdf.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14375–14385. IEEE, 2024.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7514–7528, 2021.

Yusuke Hirota, Boyi Li, Ryo Hachiuma, Yueh-Hua Wu, Boris Ivanovic, Marco Pavone, Yejin Choi, Yu-Chiang Frank Wang, Yuta Nakashima, and Chao-Han Huck Yang. Lotus: A leaderboard for detailed image captioning from quality to societal bias and user preferences. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pp. 295–309, 2025.

Hang Hua, Qing Liu, Lingzhi Zhang, Jing Shi, Soo Ye Kim, Zhifei Zhang, Yilin Wang, Jianming Zhang, Zhe Lin, and Jiebo Luo. Finecaption: Compositional image captioning focusing on wherever you want at any granularity. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24763–24773. IEEE, 2025.

Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4565–4574, 2016.

Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3128–3137, 2015.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 3992–4003. IEEE, 2023.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PmLR, 2023.

Tianle Li, Xuyang Shen, Yan Ma, Rongxin Guo, Shaoxiang Chen, Jiacheng Chen, Haochen Wang, Hongyang Tang, Yucong Zhou, and Yu Cheng. Claimdiff-rl: Fine-grained caption reinforcement learning through visual claim comparison. *arXiv preprint arXiv:2605.20278*, 2026.

Long Lian, Yifan Ding, Yunhao Ge, Sifei Liu, Hanzi Mao, Boyi Li, Marco Pavone, Ming-Yu Liu, Trevor Darrell, Adam Yala, et al. Describe anything: Detailed localized image and video captioning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21766–21777. IEEE, 2025.

Kun-Yu Lin, Hongjun Wang, Weining Ren, and Kai Han. Panoptic captioning: An equivalence bridge for image and text. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 78553–78584. Curran Associates, Inc., 2025. doi: 10.52202/085713-2631. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/7116cda41d75d580bae15d9e484a8466-Paper-Conference.pdf.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

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*, pp. 216–233. Springer, 2024a.

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, 2024b.

Zhihang Liu, Chen-Wei Xie, Bin Wen, Feiwu Yu, JixuanChen, Pandeng Li, Boqiang Zhang, Nianzu Yang, YingluLi, Zuan Gao, Yun Zheng, and Hongtao Xie. Capability: A comprehensive visual caption benchmark for evaluating both correctness and thoroughness. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc., 2025. doi: 10.52202/085713-2211. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/5f2809607f692d79a01c05c43d702883-Paper-Datasets_and_Benchmarks_Track.pdf.

Fan Lu, Wei Wu, Kecheng Zheng, Shuailei Ma, Biao Gong, Jiawei Liu, Wei Zhai, Yang Cao, Yujun Shen, and Zheng-Jun Zha. Benchmarking large vision-language models via directed scene graph for comprehensive image captioning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19618–19627. IEEE, 2025.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. ChartQA: A benchmark for question answering about charts with visual and logical reasoning. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio (eds.), *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.177. URL https://aclanthology.org/2022.findings-acl.177/.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pp. 311–318, 2002.

Runqi Qiao, Qiuna qiuna, Peiqing Yang, Yanzi Wang, Xiaowan Wang, Enhui Wan, Guanting Dong, Shiqiang Lang, Sitong Zhou, Yida Xu, Yuchen Zeng, Jie Wang, Chong Sun, Chen Li, and Honggang Zhang. We-math 2.0: A versatile mathbook system for incentivizing visual mathematical reasoning. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 150171–150218, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/f3064f7a0ca2328ecb41a3aef6177d68-Paper-Conference.pdf.

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.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 1179–1195. IEEE, 2017.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 4035–4045, 2018.

Noam Rotstein, David Bensaïd, Shaked Brody, Roy Ganz, and Ron Kimmel. Fusecap: Leveraging large language models for enriched fused image captions. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 5677–5688. IEEE, 2024.

Kuniaki Saito, Donghyun Kim, Kwanyong Park, Atsushi Hashimoto, and Yoshitaka Ushiku. Captionsmiths: Flexibly controlling language pattern in image captioning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19872–19881. IEEE, 2025.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022.

Bytedance Seed. Seed2.0 model card: Towards intelligence frontier for real-world complexity. *arXiv preprint arXiv:2607.00248*, 2026.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Aarti Sharma, Hrishikesh Singh, and Millie Pant. Pixels to prose: A comprehensive survey of image captioning techniques with deep learning and generative artificial intelligence. *Neurocomputing*, 667:132385, 2026. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2025.132385. URL https://www.sciencedirect.com/science/article/pii/S0925231225030577.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Hrishikesh Singh, Aarti Sharma, and Millie Pant. Pixels to prose: Understanding the art of image captioning. *arXiv preprint arXiv:2408.15714*, 2024.

Zhijiang Tang, Linhua Wang, Jiaxin Qi, Weihao Jiang, Peng Hou, Anxiang Zeng, and Jianqiang Huang. Cccaption: Dual-reward reinforcement learning for complete and correct image captioning. *arXiv preprint arXiv:2602.21655*, 2026.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026. URL https://qwen.ai/blog?id=qwen3.5.

Tony Cheng Tong, Sirui He, Zhiwen Shao, and Dit-Yan Yeung. G-veval: A versatile metric for evaluating image and video captions using gpt-4o. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 7419–7427, 2025.

Pavan Kumar Anasosalu Vasu, Fartash Faghri, Chun-Liang Li, Cem Koc, Nate True, Albert Antony, Gokula Santhanam, James Gabriel, Peter Grasch, Oncel Tuzel, et al. Fastvlm: Efficient vision encoding for vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19769–19780, 2025.

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*, pp. 4566–4575, 2015.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3156–3164, 2015.

Yibin Wang, Yuhang Zang, Hao Li, Cheng Jin, and Jiaqi Wang. Unified reward model for multimodal understanding and generation. *arXiv preprint arXiv:2503.05236*, 2025.

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.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng-ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

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. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 13066–13093, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/15f1dbc086bfd94d8c32557b573cbe18-Paper-Conference.pdf.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pp. 2048–2057. PMLR, 2015.

Shijia Yang, Yunong Liu, Bohan Zhai, Ximeng Sun, Zicheng Liu, Emad Barsoum, Manling Li, and Chenfeng Xu. Captionqa: Is your caption as useful as the image itself? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23741–23750, 2026.

Lin Zhang, Xianfang Zeng, Kangcong Li, Gang Yu, and Tao Chen. Sc-captioner: Improving image captioning with self-correction by reinforcement learning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23145–23155. IEEE, 2025.

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*, pp. 169–186. Springer, 2024.

A Appendix

A.1 Appendix Overview

The organization follows the structure of the DQ-RL pipeline: we first summarize the model and data configuration, then examine the evolution of the training signals. We next present qualitative comparisons and additional DQ-RL case studies, followed by the distribution of the training data and a case-level diagnostic analysis using DQ-Bench. Finally, we provide the prompt templates used throughout the training and evaluation pipeline.

A.2 Model and Data Configuration

The DQ-RL pipeline combines several models with different responsibilities. The policy model generates captions, completeness-query modules provide image-level coverage supervision, and factual-query modules decompose and verify claims made by the current rollout. The final two rows identify the models used for the DQ-Bench evaluator and the PRISM answerer. Listing these roles explicitly is important because the training policy, reward models, and evaluation models are not interchangeable.

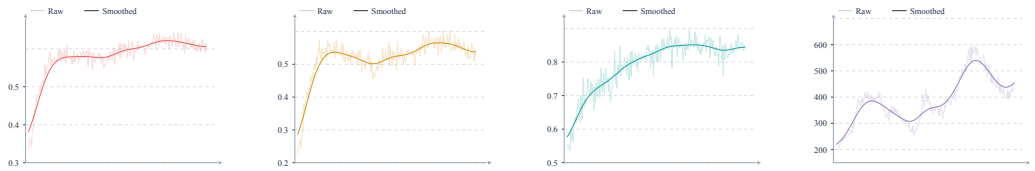
Table 7: Models and data sources used in DQ-RL training and evaluation. Each entry is associated with the role described in the surrounding text.

Component	Model / Data Source
Captioning policy initialization	Qwen3.5-4B
General CQs	CapRL data
Hard CQ generation	Seed-2.0-Lite
CQ answering	Qwen2.5-4B
FQ decomposition	Qwen3.5-4B
FQ verification	Qwen3.5-2B
DQ-Bench evaluator	Gemini-3.1-Pro
PRISM answerer	CapRL-Eval-3B

A.3 Training Dynamics

The reward design separates the optimization signal into complementary components. The overall reward summarizes the training objective, while the QA and sub-caption rewards probe different aspects of caption quality. We also track response length because a reward can be exploited by producing either unnecessarily verbose or excessively short outputs. Figure 3 reports these signals jointly, allowing us to distinguish genuine reward improvement from a length-only change.

The four curves should be interpreted together rather than independently. An increase in overall reward is most informative when it is accompanied by improvements in the query-specific rewards without an uncontrolled increase in response length. The response-length panel therefore serves as a basic training-dynamics check: it indicates whether optimization changes the amount of text while the query-based signals continue to improve.



(a) Overall reward. **(b) CQs reward.** **(c) FQs reward.** **(d) Response length.**

Figure 3: Training dynamics of DQ-RL. The panels show the evolution of the overall reward, CQ reward, FQ reward, and response length.

A.4 Qualitative Examples

The main qualitative comparison is designed to separate two common failure modes of dense image captioning: omission of visible details and the introduction of incorrect or unsupported content. Figure 4 compares the base Qwen3.5-4B model, CapRL-4B, the larger Qwen3.5-27B model, and DQ-RL on the same images. The examples are selected so that all models receive the same visual

Image	Qwen3.5-4B (Base)	CapRL-4B	Qwen3.5-27B	DQ-RL (Ours)
Breakfast still life  Visible: toast, a floral cup, translucent eyeglass lenses, and a fork and knife on the napkin.	**Omission + error** ... At the center is a hot drink in a floral-patterned cup and saucer. Its surface is covered with fine foam and is light brown. ... At the upper right is a white plate with two slices of toast, one topped with green avocado spread and garnished with green herbs. Below the plate is a folded white napkin, with a silver fork beside it. ... [Knife omitted.]	**Incorrect attributes** ... On a red round tabletop, slightly left of center, is brown foamy coffee in a floral-patterned porcelain cup. ... To the right of the coffee is a white plate holding two slices of toast, one spread with green avocado and garnished with cilantro, and the other caramel-colored. ... In the upper-left corner is a pair of tortoiseshell-framed glasses with dark lenses. ...	**Correct details** ... At the center of the foreground is a white ceramic cup and saucer decorated with pink flowers and yellow leaves. ... Behind the glasses, a silver fork and knife wrapped in a white napkin lie horizontally. At the top of the image is a white round porcelain plate holding two slices of golden toast, one piled with green avocado spread and garnished with a little green herb. Part of a silver-gray laptop keyboard with black keys is visible at the right edge. ...	**Recovered details** ... To the left of the cup, a pair of tortoiseshell-framed glasses lies diagonally, with the temples pointing left and transparent, reflective lenses. Behind the glasses is a rolled white napkin, with a silver fork and a knife lying horizontally on it and the knife handle pointing right. Above them in the background is a white round plate containing two triangular slices of toast with golden-brown marks. ...
Street-art mural  Visible: seated figure, blue-left/red-right wall, raised left hand, diamond motif, and white wall text.	**Omission + error** ... The image shows a woman seated beside a wall, with a red brick wall and a blue wall behind her, carrying several graffiti marks and pieces of text. She wears a green coat, a red inner top, blue trousers, and brown shoes, and holds a yellow cup. ... On the wall is a speech bubble with a blue diamond design, beside white text reading "phruse". ... [Left/right wall layout omitted.]	**Text/quality errors** ... Beside her is a speech bubble with a blue diamond design, with the word "PURSE" and other graffiti in the background. ... To her right is a white speech bubble containing a blue diamond design, with black graffiti letters "M" and "P" above it. ... The figure has a cartoon-like cutout style, with no realistic photographic feel, no border or watermark, no obvious focal-length or lighting effects, and image quality characterized as street-mural texture. ...	**Error + omission** ... Her right hand holds a yellow cup and her left hand hangs naturally; her expression is lively, with her mouth slightly open. On the red brick wall above the figure is a white rectangular frame containing a diamond drawn in blue lines, with white graffiti text reading "phrase" below the diamond. ... The background is textured brickwork, blue on the left and red on the right, with signs of age and stains. ... [Seated pose omitted.]	**Pose detail; OCR error** ... A street-graffiti mural features a seated female cartoon character beside a wall. ... The left side of the image is a light blue painted brick wall, and the right side is exposed reddish-brown brickwork, with a vertical boundary between them. ... Her right hand holds a yellow cup, and her left hand extends naturally with slightly spread fingers, as though conversing or presenting something. ... Below the bubble is white spray-painted text reading "phruse" ...

Figure 4: Qualitative comparison of captions generated by Qwen3.5-4B (Base), CapRL-4B, Qwen3.5-27B, and DQ-RL. The examples illustrate two complementary failure modes: omission of visible details and incorrect or unsupported attributes. Ellipses indicate omitted portions of a caption. Gray annotations mark missing details, red highlights indicate factual errors, and green highlights denote visually grounded details; in the DQ-RL column, they highlight details recovered by DQ-RL.

input, making the differences attributable to the captioning and training procedures rather than to different examples.

The base model often identifies the main subject but leaves out secondary objects, spatial relations, or fine-grained details. CapRL-4B recovers some additional content but also introduces incorrect attributes and textual errors in the examples shown here. The larger Qwen3.5-27B model provides a stronger reference point, but it can still omit details that are explicitly recovered by DQ-RL. DQ-RL combines broader visual coverage with grounded descriptions: gray annotations mark omitted details, red highlights indicate incorrect or unsupported claims, and green highlights denote visually grounded details. In the DQ-RL column, the green highlights emphasize details recovered by DQ-RL.

A.5 Additional DQ-RL Case Studies

The comparison above focuses on short excerpts so that the failure modes can be aligned across models. We next show complete DQ-RL outputs to examine how the model organizes information

DQ-RL (Ours):



Overview. A street-graffiti mural features a seated female cartoon character beside a wall. The background is blue-and-red brickwork, with a speech bubble containing a blue diamond design and white graffiti text on the right. The overall street-art style combines realism and cartoons, with vivid colors and clear lines. The image is a frontal eye-level medium close-up, with natural light and rich detail.

Details. The left side of the image is a light blue painted brick wall, and the right side is exposed reddish-brown brickwork, with a vertical boundary between them. A female figure is painted on the blue wall on the left, seated with her body turned slightly right, her head turned right, and her gaze directed toward the speech bubble on the right. She has medium-length brown hair and an expression of slight surprise or anticipation, with wide eyes and a slightly open mouth. She wears a green coat over a red top, blue jeans, and brown flat shoes. Her right hand holds a yellow cup, and her left hand extends naturally with slightly spread fingers, as though conversing or presenting something. The figure has black outlines and mostly flat colors, with some shading and dimensionality. On the red brick wall to the right is a white rectangular speech bubble containing a diamond outlined in pale blue lines, with a geometric faceted structure. Above the bubble is rough black scrawled graffiti resembling an abstract combination of the letters "M" and "D", with bold strokes. Below the bubble is white spray-painted text reading "phruse", in a bold sans-serif style, with some letters worn or distressed. Part of some pink graffiti is visible at the right edge. The overall style combines street graffiti and illustration, with relatively high color saturation, clear lines, and a poster- or sticker-like texture. Light and shadow are soft, with no obvious strong light-source direction and even illumination overall. The wall has slight texture and stains that enhance realism.

Figure 5: Complete DQ-RL caption for a street-art mural. The description first summarizes the global scene and then expands to the subject's pose, clothing, held objects, wall structure, graffiti, visible text, color palette, and lighting. This example illustrates how DQ-RL organizes a heterogeneous image into a coherent caption while retaining image-grounded details.

over a longer description. Each output uses an overview followed by fine-grained details, allowing us to inspect both global scene recognition and local visual grounding.

The first case contains a street-art mural with a seated illustrated person, multiple brick surfaces, a speech bubble, graffiti, visible text, and several small object-level attributes. This case tests whether the caption covers heterogeneous visual elements without replacing directly visible content with unsupported interpretation. Figure 5 shows that DQ-RL first establishes the scene layout and then describes the subject, pose, clothing, objects, text, colors, and rendering style.

The second case is a town panorama viewed from an elevated position. Unlike the mural example, its main challenge is not a single salient object but the coordination of foreground, middle ground, and distant scenery. As shown in Figure 6, DQ-RL explicitly describes the spatial layers, building materials, roof shapes, vegetation, hills, viewpoint, and lighting. This organization is important for dense captioning because a description that mentions only the central town would miss the depth structure and the relations between nearby and distant regions.

DQ-RL (Ours):



Overview. A European town panorama viewed from above has dense reddish-brown roofed buildings in the foreground, the extending town and a distant valley in the middle ground, and rolling green hills and pale blue sky in the background. The overall style is wide-angle landscape photography from a high downward viewpoint, with bright soft light, natural realistic colors, distinct depth layers, and a historic, tranquil atmosphere.

Details. Part of a building with a dark brown sloping roof is visible on the right in the foreground. The roof has several small arched windows or vents and appears to be made of tiles or sheet metal. Green tree branches partly obscure the view on the left. The middle ground consists of densely arranged European-style houses, mostly with dark reddish-brown or dark gray tiled roofs. Some buildings have pointed roofs or towers, including a taller pointed church tower slightly left of center, with a dark roof and a spire at the top. Most houses are low-rise, with beige, pale yellow, or gray-white walls; windows and chimneys are visible on some walls. Streets and roads can be faintly seen, with a few vehicles or road lines scattered through the scene. The middle-to-distant area is an extension of the town, where building density gradually decreases. Broad green fields and rolling hills appear in the distance. The hillside on the right is covered in dense dark green woodland, while the left hillside is gentler, with vegetation in varying shades of green. The sky is pale blue with sparse white clouds. Light comes from the upper left, casting shadows on the right sides and behind buildings. Overall illumination is soft, shadow directions are consistent, and the tone resembles natural daylight. Image clarity is high and details are rich, with a deep depth of field: foreground buildings are sharp and distant scenery slightly softer.

Figure 6: Complete DQ-RL caption for a town panorama. The description separates foreground buildings from middle-ground structures and distant hills, while also recording roof geometry, vegetation, viewpoint, lighting, and depth-dependent sharpness. The example demonstrates the model's ability to preserve scene-level spatial organization in a long caption.

A.6 Case-Level DQ-Bench Diagnostic

To make the fine-grained evaluation protocol concrete, we present a case-level analysis of a caption generated by the base Qwen3.5-4B model. The example is evaluated over the 14 atomic fields defined by DQ-Bench, covering six categories: subject, quantity, color, composition, background, and visual style. Each applicable field receives one of three labels: *Correct*, *Omit*, or *Wrong*. Fields that are not present or not applicable to the image are marked as *N/A* and excluded from the evaluation statistics.



Qwen3.5-4B (base) caption.

Overview: A red tabletop holds a hot drink, a pair of glasses, a plate of toast and avocado spread, and a laptop. The image presents a casual breakfast scene, photographed from above, with vivid colors and an everyday-life style.

Details: At the center is a hot drink in a floral-patterned cup and saucer. Its surface is covered with fine foam and is light brown. On the left is a pair of glasses with dark frames, transparent lenses, and temples pointing downward. At the upper right is a white plate with two slices of toast, one topped with green avocado spread and garnished with green herbs. Below the plate is a folded white napkin, with a silver fork beside it. Part of a white laptop keyboard and screen bezel is visible at the right edge. The background is a vivid red tabletop with a smooth surface reflecting soft light. The composition is compact, with strong color contrasts and a relaxed, pleasant atmosphere suited to an everyday leisure setting.

Atomic field	Label	Visual evidence and rationale
Subject / Main subject	Correct	Lists the drink, glasses, toast plate, and laptop.
Subject / Facial features	N/A	No person or face appears in this still life.
Subject / Action	N/A	There is no visible subject action to evaluate.
Subject / Pose	N/A	No person or animal pose is present.
Subject / Expression	N/A	No face or expression is present.
Quantity / Count	Wrong	Says two toast pieces; the reference counts three.
Color / Color	Correct	Red table, green avocado, white plate, and brown drink match.
Composition / Overall layout	Correct	Cup central, plate above, and laptop at right match.
Composition / Spatial relation	Correct	Glasses left of cup and plate above it are located correctly.
Background / Scene	Omit	The patterned upper-left edge outside the table is absent.
Background / Position and coverage	Omit	Does not say the red table fills most of the frame.
Visual style / Photographic	Correct	Identifies an everyday-life photograph.
Visual style / Color style	Correct	Vivid red–green–white contrast is described.
Visual style / Artistic/visual	Correct	The casual breakfast still-life style is appropriate.

This example illustrates how DQ-Bench separates different types of caption errors. The base caption correctly describes the main objects, colors, layout, spatial relations, and photographic style, but it reports an incorrect toast count and omits parts of the surrounding scene and frame coverage. In total, the example contains seven Correct fields, two Omit fields, one Wrong field, and four N/A fields. The case demonstrates that the protocol distinguishes missing information from factually incorrect content rather than collapsing both into a single caption-level score.

A.7 Data and Question Distributions

The training data are constructed to provide both broad visual coverage and question-level supervision. Figure 7(a) shows the hierarchical image-category distribution. The inner ring summarizes the four major domains—Nature, Design, People, and Synthetic—while the outer ring gives a finer-grained breakdown of the corresponding subcategories. This hierarchy exposes the diversity of the training images rather than reporting only a single aggregate count.

Figure 7(b) reports the number of questions associated with each image. Most images are paired with five to seven questions, with six questions being the most frequent setting. The remaining images contain fewer or more questions, reflecting the adaptive construction of queries for different image contents. This distribution is relevant to DQ-RL because it determines how many complementary visual facts can contribute to the completeness reward for a given image.



(a) Image category distribution. **(b) Question count distribution.**

Figure 7: Distribution of the training images and associated questions. The left panel shows the hierarchical image taxonomy, and the right panel shows the frequency of images grouped by the number of questions assigned to each image.

A.8 PROMPT TEMPLATES

The prompts below specify how visual information is converted into the different supervision signals used by DQ-RL. They are included because the prompt wording determines what counts as an observable detail, how difficult queries are constructed, and how caption claims are scored. Each prompt is preceded by a short description of its role in the pipeline.

Caption Generation. The caption-generation prompt asks the policy or captioning model to describe directly observable content, including subjects, counts, attributes, spatial relations, text, style, viewpoint, and lighting. This prompt defines the target form of dense captioning and discourages unsupported inference.

Prompt for Image Caption Generation

````
<image> You are a professional image-to-text prompt generator.
Based only on the given image, describe directly observable
content. Your description should cover the main subjects and their
counts; colors; actions and poses; spatial relations;
interactions; foreground and background; original visible text and
its layout (if any); visual and artistic style; color palette;
photographic style; lens and framing; shooting angle; lighting;
image quality; and salient elements such as borders, watermarks,
logos, or screenshot interfaces. For a close-up of a person or
animal, also describe facial features and expression. Do not infer
invisible information, add content that is not in the image,
translate visible text, or provide an explanation. Return the
result directly in the following YAML format. The overview should
be one sentence summarizing the core subjects, their counts and
relations, background, style, framing, and viewpoint.
Return the following YAML structure:
```yaml
Overview: |
  Overview content...
Details: |
  Detailed content...
```
````

Hard Completeness Query Construction. Given an image and an initial caption, this prompt generates targeted questions about details that are missing, underspecified, or potentially misstated. The resulting hard CQs focus the completeness signal on the current policy's weaknesses rather than sampling only easy facts.

Hard CQ Generation

```
[System]
You are a caption-error QA generator. Given an image and a
complete caption, generate image-grounded multiple-choice
questions targeting important omissions or unsupported claims.
Use only visible image evidence to determine correct answers. Do
not use common sense, world knowledge, or the caption itself as
evidence. Consider both overview and details: information
mentioned anywhere is covered. Preserve subject identity and
attribute associations.
Cover subjects/counts, colors, actions/poses, spatial relations,
interactions, foreground/background, visible text/layout, visual
and photographic style, framing/viewpoint, lighting, image
quality, borders/watermarks/logos/UI, and close-up facial
features/expression.
Each question must target one genuine mismatch. Do not mention the
caption or its errors in questions. Provide four options A/B/C/D
with plausible distractors and one image-supported correct answer.
For visually unverifiable claims, an option may state that the
image provides no evidence; do not use it to evade a visible
contradiction.
Do not rewrite the caption or generate questions about formatting,
word count, prose style, or unrelated content.
Return JSON only:
{"qas": [{
  "question": "string",
  "options": {"A": "string", "B": "string", "C": "string", "D":
  "string"},
  "answer": "A|B|C|D"
}]}
[User]
Image: <image>
Caption: {caption}
Compare the image with the entire caption. Generate up to 10
nonduplicate questions about important visible details omitted or
incorrectly described. Return fewer when fewer genuine mismatches
exist; never invent errors to reach 10.
```

Completeness Query Answering. This prompt evaluates whether a caption contains enough information for a text-only answerer to answer an image-grounded completeness question. Its output contributes to the coverage-oriented reward and therefore measures whether the caption exposes relevant visual facts in language.

Prompt for CQ Answering

```
[User]
You will be given an image caption describing the visual content.
Your task is to answer the multiple-choice question strictly based
on the caption, even if the answer may seem obvious from prior
knowledge or question wording.
Ignore any external knowledge. Do not make assumptions beyond what
the caption explicitly or implicitly states.
Example 1:
Caption: <Caption Start> A woman in a red coat is walking a black
dog across a snowy park. <Caption End>
Question: What color is the dog?
- A) Brown
- B) White
- C) Black
- D) Gray
- E) Can not answer based on the caption
The answer is C.
Example 2:
Caption: <Caption Start> A child is waving a British flag during a
parade.
<Caption End>
Question: What color is the flag?
- A) Red
- B) Blue
- C) Red, white, and blue
- D) White
- E) Can not answer based on the caption
The answer is E.
Now, answer the question based on the following caption:
Caption: <Caption Start> {caption} <Caption End>
Question: {question}
[Assistant prefix]
The answer is
```

Faithfulness Query Decomposition. This prompt decomposes a rollout caption into semantic factual statements. The decomposition creates a set of verifiable claims that is more diagnostic than assigning one holistic score to the entire caption, especially when most of the caption is correct but one local attribute is hallucinated.

FQ Decomposition

```
[System]
Follow the user's instructions exactly. Return valid JSON only,
without Markdown fences or commentary.
[User]
Decompose the complete caption, including both Overview and
Details when present, into FQs at semantic-statement granularity.
Rules:
- Each FQ must be a self-contained visual factual claim with
enough original subject information to distinguish similar people
or objects. Use clothing, color, position, action, relation,
visible text, carried objects, or scene region as anchors. Avoid
ambiguous references such as he, she, it, this person, the man, or
the woman when multiple similar subjects may exist.
- Merge repeated or semantically equivalent claims. Keep related
details in one naturally verifiable unit: clothing/color;
action/key object or spatial relation; pose/body-part details;
text/location; UI module/value; chart module/nodes/arrows;
structure/components; background/lighting/atmosphere.
- Do not split one action, pose, structure, text block, UI/chart
module, or scene region merely because it contains different parts
or sub-elements. Unrelated facts about different subjects,
regions, or dimensions may split.
- Retain all explicitly stated facts, including exact text and
numbers. Do not add, delete, correct, translate, summarize, or
infer information.
- Match the caption's primary language. Prefer compact claims;
3--8 is typical for a detailed caption, but use its actual content
rather than a fixed count.
Example:
Too fine-grained: ["The runner bends an arm.", "One leg of the
runner extends forward.", "The runner leans forward."]
Better: ["The runner leans forward with bent arms and one leg
extended forward in a running pose."]
Return only a JSON string list, not an object or fields such as
entity, entity_anchor, location, detail, entity_id, or
entity_type.
Caption: {caption}
```

Faithfulness Query Scoring. This prompt checks the decomposed statements against the visual evidence and assigns scores based on whether the claims are grounded, omitted, or wrong. The resulting signal complements completeness queries by penalizing content that is fluent but unsupported by the image.

FQ Scoring

```
[System]
Judge only clearly visible image content. Do not fill in
information from common sense, the original caption, or guesses.
Return valid JSON only.
[User]
Judge the FQ against the image. Return {"score": value} without
explanation.Return {"score": null} only when the image itself
genuinely prevents judgment.
Scoring rules:
1.00: The core fact and all important details are supported by the
image.
0.75: The core fact is correct, with only minor inaccuracies in
ancillary details.
0.50: The core fact is correct, but important ancillary details
are incorrect or incomplete.
0.25: Some facts are correct, but the core action, core attribute,
or spatial relation has a clear error.
0.00: The core fact is wrong, including incorrect subject
identity, a nonexistent core object, incorrect count, incorrect
main action, incorrect key spatial relation, or fabricated content
absent from the image.
If the core fact is wrong, the maximum score is 0.25.
If the core fact is correct but important details are wrong, the
maximum is 0.5. Do not penalize omission of other image facts;
judge only this FQ's content. If the image supports a specific
attribute but the FQ says "A or B" or "possibly A," treat the
ambiguous expression as incorrect.
Image: <image>
FQ: {FQ}
```