

# CAPRL: STIMULATING DENSE IMAGE CAPTION CAPABILITIES VIA REINFORCEMENT LEARNING

Anonymous authors

Paper under double-blind review

## ABSTRACT

Image captioning is a fundamental task that bridges the visual and linguistic domains, playing a critical role in pre-training Large Vision-Language Models (LVLMs). Current state-of-the-art captioning models are typically trained with Supervised Fine-Tuning (SFT), a paradigm that relies on expensive, non-scalable data annotated by humans or proprietary models. This approach often leads to models that memorize specific ground-truth answers, limiting their generality and ability to generate diverse, creative descriptions. To overcome the limitation of SFT, we propose applying the Reinforcement Learning with Verifiable Rewards (RLVR) paradigm to the open-ended task of image captioning. A primary challenge, however, is designing an objective reward function for the inherently subjective nature of what constitutes a “good” caption. We introduce Captioning Reinforcement Learning (CapRL), a novel training framework that redefines caption quality through its utility: a high-quality caption should enable a non-visual language model to accurately answer questions about the corresponding image. CapRL employs a decoupled two-stage pipeline where an LVLM generates a caption, and the objective reward is derived from the accuracy of a separate, vision-free LLM answering Multiple-Choice Questions based solely on that caption. As the first study to apply RLVR to the subjective image captioning task, we demonstrate that CapRL significantly enhances multiple settings. Pretraining on the CapRL-5M caption dataset annotated by CapRL-3B results in substantial gains across 12 benchmarks. Moreover, within the Prism Framework for caption quality evaluation, CapRL achieves performance comparable to Qwen2.5-VL-72B, while exceeding the baseline by an average margin of 8.4%. Results validate that our CapRL effectively trains models to produce more general and accurate image descriptions, moving beyond the limitations of traditional SFT-based image captioning models.

## 1 INTRODUCTION

The image captioning task (Karpathy & Fei-Fei, 2015; Vinyals et al., 2015), which generates a natural language description for a given image, bridges the gap between the visual and linguistic worlds. The captioning capability is fundamental to various applications, including vision-language models like CLIP (Radford et al., 2021), which learn a shared embedding space for images and text. Furthermore, captions are often a core component in the pre-training stage of Large Vision-Language Models (LVLMs) (Liu et al., 2023), where the model learns to align visual information with linguistic descriptions on a massive scale before being fine-tuned for other downstream tasks.

Given the importance of image captioning, there is a strong need for captioning models that can provide dense and accurate descriptions. Most modern captioning models Chen et al. (2024b); Rotstein et al. (2024); Vasu et al. (2025) are trained based on LVLMs using Supervised Fine-Tuning (SFT). While effective, SFT requires large datasets annotated by humans or proprietary models, which are **expensive** and **not scalable**. Furthermore, image captioning is an inherently open-ended problem, where a single image can be accurately described by a wide variety of captions. Since SFT models are trained to match a single ground-truth description for each image, they tend to **memorize specific answers** rather than learning the underlying concepts. As a result, the SFT models become **less general** and struggle to generate the diverse range of valid captions possible for a single image.

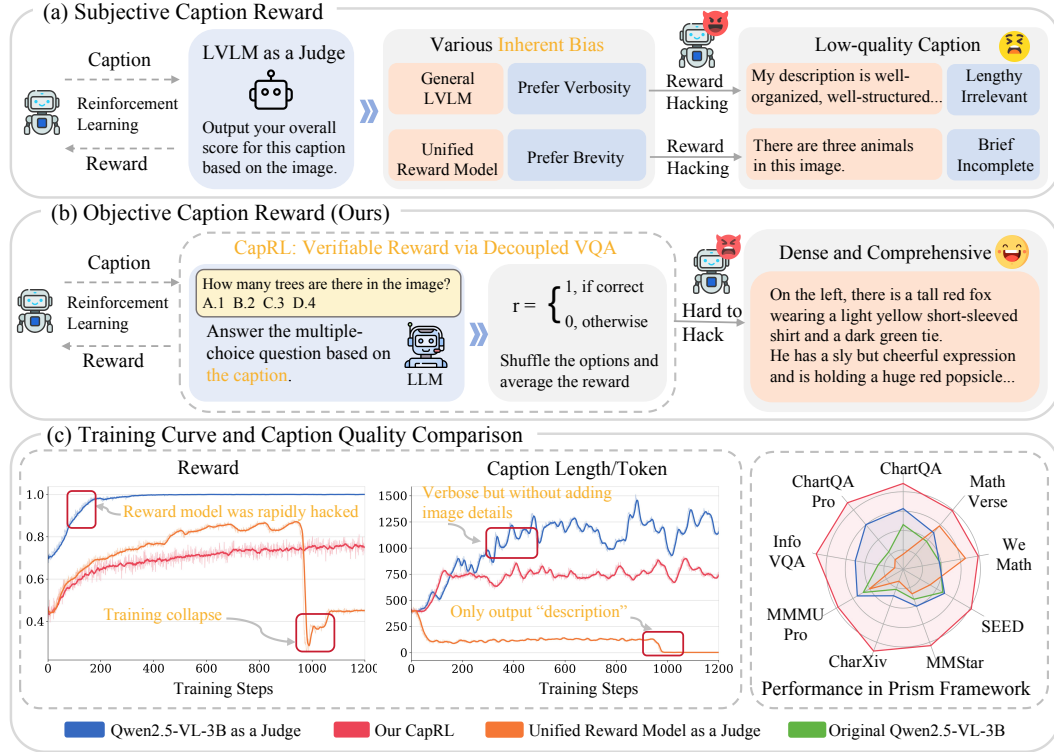


Figure 1: (a) **Existing Reward Models:** Current LVLM-as-a-judge/reward models suffer from limitations like rewarding verbosity or brevity, leading to low-quality captions and reward hacking. (b) **Our CapRL:** CapRL uses a decoupled two-stage VQA approach to provide subjective rewards for captions. (c) **CapRL’s Advantage:** CapRL outperforms previous subjective reward methods, as shown by training curves and higher performance in the Prism (Qiao et al., 2024) evaluation setting.

The limitations of SFT have led to a recent paradigm shift in the post-training of LVLMs toward Reinforcement Learning with Verifiable Rewards (RLVR) (Lambert et al., 2024). RLVR is the paradigm that trains models by providing clear and objective reward from the verifier, such as a binary signal of correctness for mathematical reasoning (e.g., DeepSeek-R1 (Guo et al., 2025)). Unlike SFT, which teaches a model to mimic a single ground-truth response, RLVR encourages the model to generate more diverse and robust outputs that meet the verifiable criteria. Our objective is to design a powerful and scalable RLVR training paradigm for the image captioning task to generate more creative and more general variety of accurate descriptions.

However, applying RLVR to open-ended tasks like image captioning is challenging, primarily due to the difficulty of designing an *objective* reward function. A good caption can be *subjective*, with multiple valid descriptions possible for the same image. **Unified Reward Model (Wang et al., 2025)** is a widely adopted model trained to provide reward signals, and we conducted experiments with this specialized model, as well as general LVLMs. As shown in Fig. 1 (a), these early studies fail to provide accurate reward signals for RL training. Using **reward models** (Liu et al., 2025a; Su et al., 2025; Lu, 2025) or **LLM-as-a-judge** (Gunjal et al., 2025) to provide feedback is vulnerable to *reward hacking*. The captioning model learns to exploit weaknesses in the reward models (e.g., verbosity or brevity outputs) rather than producing a high-quality response. Moreover, it is difficult to create effective rubrics or evaluation prompts for LVLM-as-a-judge methods because captions are free-form and encode substantial information. Using **reference answer** as rewards (Gurung & Lapata, 2025; Yu et al., 2025) like ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002) is constrained when evaluating complex and long-form captions. Fig. 1 (c) further demonstrates the limitations of previous subjective caption rewards, showing reward hacking and unstable training curves.

To design the *objective* RLVR reward function for the *subjective* image captioning task, we introduce a novel perspective, where a caption’s quality is proportional to its utility. When the image caption is detailed and accurate, a text-based LLM that can’t directly “see” the image can still answer Visual Question Answering (VQA) questions about the image. For example, for the question “What

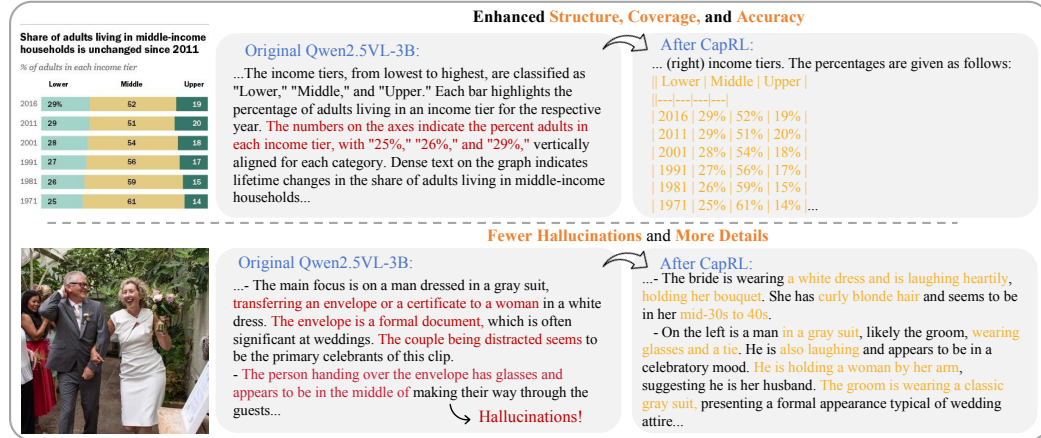


Figure 2: Illustration of the captioning capability improvement CapRL brings to Qwen2.5-VL-3B.

color is the frisbee?”, the LLM finds the phrase “red frisbee” in the caption and correctly answers “red.” Driven by this motivation, we present an effective decoupled two-stage pipeline, dubbed as **Captioning Reinforcement Learning (CapRL)**, as shown in Fig. 1 (b). Specifically, the reward of our CapRL framework is determined by how well a caption generated by an LVLM enables a separate non-visual LLM to answer Multiple-Choice Questions (MCQs) about the source image. The LLM’s resulting accuracy serves as the objective reward for the RLVR training. To ensure the high-quality MCQs data that present enough knowledge required for VQA has been examined, we also developed a specific QA curation pipeline. The images are sampled from various sources, including natural images, charts, and documents. The questions and answers are filtered to ensure the questions can only be answered by analyzing the image content itself.

We conduct a comprehensive evaluation of the significant benefits brought by CapRL. From a qualitative perspective, as shown in Fig. 2, applying the CapRL framework to Qwen2.5-VL-3B makes its outputs more well-organized and accurate. Further illustrative cases for various charts, infographics, or natural images can be found in Appendix A. From a quantitative perspective: (i) We employ CapRL-3B to annotate the CapRL-5M caption dataset, and LVLM pretraining on this dataset yields substantial improvements across 12 benchmarks. (ii) Furthermore, using the Prism Framework (Qiao et al., 2024) for caption quality evaluation, we observed that CapRL-3B remarkably achieves performance comparable to the 72B model, and outperforms the baseline by an average margin of 8.4%. These results demonstrate that our CapRL framework, by leveraging objective reward design as a reliable optimization signal, effectively drives the model to produce dense and accurate captions.

Our contributions are summarized as follows:

- 1) We contribute the first study of applying Reinforcement Learning with Verifiable Rewards for the open-ended and subjective image captioning task. Unlike traditional Supervised Fine-Tuning, which can lead to models memorizing a limited set of annotated captions, our method allows the model to explore and generate a broader range of creative and general descriptions.
- 2) We present CapRL, a new training paradigm featuring a decoupled two-stage pipeline. The initial stage uses LVLMs to generate rich and accurate captions. Subsequently, the second stage evaluates caption quality by using a vision-free LLM to perform the QA task. We also created a specific QA curation pipeline to ensure the quality of the questions and answers used for the second stage.
- 3) We carry out extensive experiments to verify the effectiveness of CapRL. Notably, both in the LVLM Pretraining setting for modality alignment and the Prism setting for caption informativeness evaluation, CapRL consistently exhibits superior performance compared to the baselines.

## 2 RELATED WORK

**Image Captioning.** Early Large-scale image-text corpora (Schuhmann et al., 2022; Changpinyo et al., 2021; Thomee et al., 2016) have driven vision-language pretraining. To scale and improve captions, researchers design advanced captioning pipelines: BLIP-LAION (Li et al., 2022) generates short synthetic captions, LaCLIP (Fan et al., 2023) uses ChatGPT to rewrite them, and CapsFusion

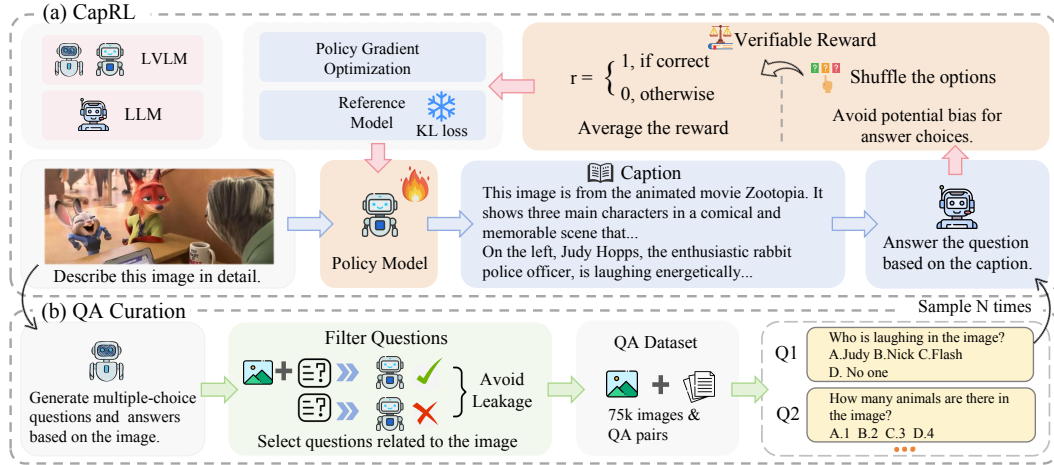


Figure 3: **Overview of CapRL.** Unlike the standard single-stage RLVR, our CapRL performs a decoupled two-stage process. Captions generated by the LVLM, paired with curated MCQs (b), are used to query an LLM, whose resulting accuracy becomes the objective reward for the LVLM (a). Our CapRL offers a scalable framework for applying RLVR to the open-ended image captioning task.

(Yu et al., 2024) consolidates and refine information with fine-tuned models. Besides, there are many research projects which use GPT-4V and human-in-the-loop pipelines to produce richer, fine-grained annotations such as ShareGPT4V (Chen et al., 2024b) and ALLaVA (Chen et al., 2024a). Recent studies (Li et al., 2024; Sun et al., 2024) have explored multi-expert approaches to compensate for LVLM limitations. In summary, some works rely on complex pipelines with multiple models, training-free but costly at inference, while others require lots of expensive labeled data for SFT. In contrast, our CapRL achieves strong performance with remarkable data efficiency through RLVR.

**Reinforcement Learning with Verifiable Rewards (RLVR).** RLVR (Lambert et al., 2024) represents a promising paradigm for training Large Language Models (LLMs) to tasks that have an objective, easily verifiable reward signal. For example, in mathematical problem-solving, the reward can be a binary signal of correctness (Shao et al., 2024), and for code generation, it can be whether the code passes unit tests (Team et al., 2025). Compared to the traditional Supervised Fine-Tuning (SFT), RLVR offers a more robust and scalable approach. While SFT trains models to imitate a set of provided ground-truth answers, often leading to models that memorize specific phrasings (Chu et al., 2025), RLVR encourages the model to explore and discover optimal solutions. This is particularly beneficial for problems with multiple valid answers or reasoning paths.

### 3 METHODOLOGY

An overview of our CapRL is shown in Fig. 3. The CapRL framework consists of a novel, decoupled two-stage process. In the first stage, an LVLM generates a caption for an input image. In the second stage, this caption, along with a series of MCQs, is provided as input to an LLM. In the following, we will describe how to apply RLVR on the image captioning task via our CapRL in Section 3.1. Then we use the model trained with CapRL to construct the CapRL-5M dataset in Section 3.2.

#### 3.1 CAPRL

The design of the reward function is a pivotal factor in the success of RLVR-based approaches, since the reward function directly guides the optimization direction of the policy model. Although designing reward functions for objective tasks (Shao et al., 2024; Liu et al., 2025b; Luo et al., 2025) is straightforward, developing the reward function for the subjective image captioning task is challenging. While reward models (Liu et al., 2025a; Su et al., 2025; Lu, 2025) or the “LLM-as-a-judge” approach (Gunjal et al., 2025) have been explored for RL training on open-ended tasks, these models are still vulnerable to exploitation in captioning task, primarily owing to their intrinsic biases, which may unintentionally encourage the captioning model to produce verbose or brief results.

To design a reliable verifiable reward module, we leverage a perception-reasoning decoupled VQA task as a proxy to evaluate the quality of captions. The overall process of our proposed method CapRL, is illustrated in Fig. 3. During the GRPO training process, an image and an instruction are first provided as input to the policy model to sample a set of candidate captions. Each caption is then paired with corresponding questions and fed to a Large Language Model (LLM). We assign each caption a reward score based on the accuracy of answers generated by the LLM. Subsequently, we calculate the mean and variance of rewards across the group to derive the advantage for each caption. To ensure training stability, and consistent with the original GRPO framework, we incorporate a KL-divergence penalty. The policy model is then updated via policy gradient optimization.

To prepare the data for GRPO training, we constructed a VQA dataset composed exclusively of multiple-choice questions. This multiple-choice format facilitates the computation of verifiable rewards. Throughout this curation process, we utilized an LVLM to filter the data and prevent data leakage. Further details regarding our reward design and QA curation are provided below.

**Reward Design.** Specifically, given an instruction and an image, the policy model  $\mathcal{M}_V$  generates a set of captions  $\{c_1, c_2, \dots, c_G\}$ . Each caption is then paired with questions related to the image and passed to a large language model (LLM), denoted as  $\mathcal{M}_L$ , for answering. Since the  $\mathcal{M}_L$  does not have access to the image directly, its ability to answer the question correctly depends entirely on how comprehensive and accurate the caption is. Captions that include more relevant objects and detailed descriptions, are more likely to provide the necessary information for the LLM to answer a question correctly. In contrast, less informative captions are more likely to lead to incorrect answers. Since LLMs exhibit high stability in answering multiple-choice questions, and the evaluation of their responses only requires exact matching, the accuracy of the LLM’s responses can therefore serve as a reliable indicator of caption quality. This question-answering process can be formulated as:

$$a_m = \mathcal{M}_L(c_i, q_m), \quad (1)$$

where  $q_m$  denotes the  $m$ -th question associated with current image  $I$ , and  $a_m$  is the LLM’s answer to that question. Then the reward for a single question is computed using a simple exact-match criterion:

$$r(a_m) = \begin{cases} 1, & \text{if } a_m = \text{GT}_m, \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

Here,  $\text{GT}_m$  is the ground-truth answer to question  $q_m$ .

To eliminate potential bias in the LLM’s preference for specific answer choices, we randomly shuffle the options each time a question is presented. Additionally, relying on a single answer to evaluate a caption lacks robustness. To ensure the stability of caption scoring, we sample  $N$  times from all the questions related to the image and let  $\mathcal{M}_L$  answer them independently. The final reward for a caption is computed as the average accuracy over these  $N$  sampled questions. Formally:

$$R_{c_i} = \frac{1}{N} \sum_{k=1}^N r(\mathcal{M}_L(c_i, \text{Shuffle}(q_{m_k}))), \quad m_k \sim \{1, \dots, M\}. \quad (3)$$

Here,  $M$  denotes the number of questions associated with the current image  $I$ . Since we compute the caption reward directly from the original caption, there is no need to perform intermediate reasoning steps as in DeepSeek-R1, which first carries out a thinking process before formatting an answer. As a result, our method avoids the need for any format-specific rewards and retains a clean, flexible reward computation process that fully respects the free-form nature of the policy model’s output. It is important to note that, in our GRPO training setup, Qwen2.5-3B-Instruct is used as  $\mathcal{M}_L$  by default, which makes the overall training highly efficient.

**QA Curation.** To train CapRL effectively, a high-quality VQA dataset  $(q, a)$  with question  $q$  and answer  $a$  is required to provide reliable reward signals. We construct this VQA dataset using a structured three-stage curation pipeline. (1) Image Collection. We begin by sourcing diverse images from the web and existing open-source datasets, including natural scenes, charts, and documents, to maximize variety. (2) QA Generation. For each image, we then use Qwen2.5-VL-72B (Bai et al., 2025) to automatically generate multiple question-answer pairs. (3) QA Filtering. Finally, we implement a stringent QA filtering process to ensure the quality of the generated QA pairs. The QA filtering stage is to verify that all questions are strictly visually-grounded and answerable exclusively through analysis of the image content. The final QA filtering stage is crucial to prevent information leakage and guarantees that the model must perform true visual understanding, rather than relying on external knowledge or cues within the question itself to answer the generated questions.

Specifically, the filtered set of QA pairs, denoted as  $\mathcal{Q}$ , is then defined as:

$$\mathcal{Q} = \{(q, a) \in \mathcal{D} \mid \mathcal{M}_{V_f}(q, I) = a \wedge \mathcal{M}_{V_f}(q) \neq a\}, \quad (4)$$

where  $(q, a)$  is a question-answer pair from the initial generated dataset  $\mathcal{D}$ ,  $I$  is the corresponding input image,  $\mathcal{M}_{V_f}$  is the LVLM used in QA Filtering,  $\mathcal{M}_{V_f}(q, I)$  represents the answer generated when conditioned on both the question  $q$  and the image  $I$ , and  $\mathcal{M}_{V_f}(q)$  is the answer generated when the image is omitted. According to Eq. (4), the QA filtering step ensures that each selected QA pair requires the image context to be answered correctly. To manage computational costs effectively, the QA filtering step is performed using the Qwen2.5-VL-3B model (Bai et al., 2025) as  $\mathcal{M}_{V_f}$ .

After filtering, we retain approximately 75k images along with their corresponding QA pairs to train the CapRL captioning model. Please refer to Appendix F and I for the curation details.

### 3.2 CAPRL-5M DATASET

By employing our carefully designed CapRL training scheme, we obtained CapRL-3B, and further used this powerful captioner to annotate 5M images, ultimately forming CapRL-5M.

**Image Collection and Processing.** In collecting images, we primarily considered diversity, quality, and safety. Among the currently high-quality open-source image datasets, ShareGPT4V-1M (Chen et al., 2024b) and DenseFusion-1M (Li et al., 2024) are relatively large in scale. Since both datasets have already undergone extensive filtering and clustering to ensure image quality, we directly incorporated all images from them. To further enhance dataset diversity, we also gathered a large number of images from the web, spanning natural photographs, documents, charts, and user interfaces. However, the quality of web images is highly uneven, and they pose potential safety risks, which could severely impact both model training and deployment safety. To address this, we applied rigorous filtering and ultimately retained 3M high-quality images. Combined with the two open-source datasets, this yielded a total of 5M images. The detailed filtering process is described in Appendix H.

**Caption Model selection.** In typical multimodal pretraining scenarios, the pretraining dataset often requires a massive number of image-text pairs, making annotation costs substantial. Considering practical applications, we decide to train a highly lightweight yet powerful captioner to keep annotation costs more acceptable. Specifically, we initialize the policy model with Qwen2.5-VL-3B and employ our CapRL framework, resulting in the powerful CapRL-3B model as the captioner.

## 4 EXPERIMENTS

### 4.1 PRETRAINING SETTING

To thoroughly evaluate the quality of the CapRL-5M dataset, we conduct comprehensive comparisons with widely used caption datasets from the open-source community.

**Implementation Details.** In our setup, the language model is initialized with a pretrained LLM, the visual encoder with a pretrained ViT, and the MLP projector randomly, following a standard multimodal pretraining scheme. We conduct experiments under three settings: Qwen2.5-3B + Qwen2.5-ViT, Qwen2.5-7B + Qwen2.5-ViT, and InternLM2.5-7B + CLIP-ViT-L. Training follows the ShareGPT4V paradigm in three stages: Initial Alignment with BLIP-558K dataset (Li et al., 2022); Further Pretraining with diverse high-quality image-caption datasets; and SFT with Open-LLaVA-NeXT-1M (Chen & Xing, 2024). For comparison, we adopt strong baselines including Vanilla, which skips Further Pretraining, ShareGPT4V-1M, DenseFusion-1M, and CapRL-1M (randomly sampled from CapRL-5M). Detailed training details are provided in Appendix J.

**Main Results.** As shown in Table 1, when using CapRL-1M as the further pretraining dataset, performance on the vast majority of benchmarks surpasses both ShareGPT4V-1M and DenseFusion-1M. Specifically, under the Qwen2.5-3B + Qwen2.5-ViT setting, it exceeds DenseFusion-1M by 6.8% on InfoVQA, and outperforms by 2.7% and 3.6% on DocVQA and ChartVQA. These remarkable results indicate that CapRL-3B is effective for domains such as documents, charts, and infographics, which demand fine-grained perception and structured description. The captions in CapRL-1M are highly detailed and accurate for such image types, enabling LVLMs to achieve better modality alignment and a deeper understanding of the corresponding visual features.



Table 1: Performance comparison using different pretraining datasets. CapRL-1M significantly outperforms other datasets across all 3 settings, and further improvements are observed when scaling the data to 5M. The best results are **bold** and the second-best results are underlined.

Pretraining Dataset	Info VQA	Doc VQA	Chart QA	Real WorldQA	Math Vista	SEED2 Plus	MME RW	MMB	MMStar	MMVet	AI2D	GQA	Average
<i>Qwen2.5-3B + Qwen2.5-ViT</i>													
Vanilla	43.9	81.0	72.7	55.1	41.6	56.6	30.5	68.6	44.7	41.0	68.3	61.5	55.5
ShareGPT4V-1M	46.1	82.4	74.2	55.0	44.7	60.5	29.8	68.9	45.2	42.4	70.1	61.4	56.7
DenseFusion-1M	49.4	84.6	74.4	54.1	44.6	59.1	<u>30.7</u>	69.0	45.6	40.2	70.4	<u>62.5</u>	57.1
CapRL-1M	<u>56.2</u>	<u>87.3</u>	<u>78.0</u>	<u>55.1</u>	<u>45.5</u>	<u>62.0</u>	30.3	<u>70.5</u>	<u>47.0</u>	<u>50.0</u>	<u>72.9</u>	61.6	<u>59.7</u>
CapRL-5M	<b>61.5</b>	<b>90.0</b>	<b>80.5</b>	<b>57.6</b>	<b>48.1</b>	<b>63.2</b>	<b>30.9</b>	<b>73.1</b>	<b>50.4</b>	<b>52.6</b>	<b>74.7</b>	<b>62.6</b>	<b>62.0</b>
<i>Qwen2.5-7B + Qwen2.5-ViT</i>													
Vanilla	47.6	83.7	77.1	55.9	47.4	60.4	29.4	72.1	48.1	47.1	72.4	62.7	58.7
ShareGPT4V-1M	49.8	85.1	75.7	56.8	46.6	60.9	31.8	71.9	48.4	45.9	72.2	62.7	59.0
DenseFusion-1M	53.5	87.8	76.7	58.6	46.3	61.0	31.1	<u>72.6</u>	48.6	49.7	72.5	63.1	60.2
CapRL-1M	<u>59.9</u>	<u>89.5</u>	<u>80.6</u>	<u>58.9</u>	<u>50.4</u>	<u>63.1</u>	<u>32.2</u>	72.1	<u>51.3</u>	<u>50.5</u>	<u>75.3</u>	<u>63.2</u>	<u>62.2</u>
CapRL-5M	<b>63.4</b>	<b>91.4</b>	<b>81.5</b>	<b>61.4</b>	<b>50.8</b>	<b>63.2</b>	<b>34.9</b>	<b>72.7</b>	<b>52.6</b>	<b>52.6</b>	<b>76.9</b>	<b>63.8</b>	<b>63.8</b>
<i>InternLM2.5-7B + CLIP-ViT-L</i>													
Vanilla	37.4	73.2	68.7	56.9	44.2	58.2	30.7	70.7	47.0	43.1	71.8	64.9	55.6
ShareGPT4V-1M	38.9	73.8	69.8	56.3	44.8	59.9	33.2	72.6	46.2	43.3	72.7	65.0	56.4
DenseFusion-1M	39.3	76.4	70.8	<b>59.7</b>	44.5	60.3	34.1	72.2	47.9	44.0	73.7	65.5	57.4
CapRL-1M	<u>43.3</u>	<u>80.0</u>	<u>75.8</u>	<u>58.0</u>	<u>49.6</u>	<u>62.8</u>	<u>34.1</u>	<u>73.4</u>	<u>50.2</u>	<u>46.6</u>	<u>76.0</u>	<u>65.8</u>	<u>59.6</u>
CapRL-5M	<b>47.0</b>	<b>83.5</b>	<b>77.7</b>	<b>59.7</b>	<b>50.4</b>	<b>63.5</b>	<b>38.9</b>	<b>73.7</b>	<b>53.3</b>	<b>54.3</b>	<b>77.6</b>	<b>66.3</b>	<b>62.2</b>

Table 2: Ablation on image sources. We annotate the images in ShareGPT4V-1M and DenseFusion-1M using CapRL-3B, and use them respectively as pretraining datasets for comparison.

Pretraining Dataset	Info VQA	Doc VQA	Chart QA	Real WorldQA	Math Vista	SEED2 Plus	MME RW	MMB	MMStar	MMVet	AI2D	GQA	Average
<i>Qwen2.5-3B + Qwen2.5-ViT</i>													
Vanilla	43.9	81.0	72.7	55.1	41.6	56.6	30.5	68.6	44.7	41.0	68.3	61.5	55.5
ShareGPT4V-1M	46.1	82.4	74.2	55.0	44.7	<b>60.5</b>	29.8	68.9	45.2	42.4	70.1	61.4	56.7
CapRL-ShareGPT4V-1M	<b>52.1</b>	<b>85.9</b>	<b>75.2</b>	<b>56.3</b>	<b>45.6</b>	60.0	<b>30.9</b>	<b>70.9</b>	<b>46.7</b>	<b>47.5</b>	<b>71.4</b>	<b>61.7</b>	<b>58.7</b>
DenseFusion-1M	49.4	84.6	74.4	54.1	44.6	59.1	30.7	69.0	45.6	40.2	70.4	<b>62.5</b>	57.1
CapRL-DenseFusion-1M	<b>55.0</b>	<b>87.8</b>	<b>77.5</b>	<b>56.2</b>	<b>44.7</b>	<b>62.8</b>	<b>32.0</b>	<b>71.0</b>	<b>46.6</b>	<b>49.9</b>	<b>72.7</b>	62.3	<b>59.9</b>

CapRL-5M further demonstrates consistently superior performance across all 12 benchmarks. These results highlight the strong scaling properties of the CapRL-3B-annotated dataset: as the training data size expands from 1M to 5M, model performance continues to improve steadily. This phenomenon underscores the practical value of CapRL for multimodal pretraining, as it enables the construction of high-quality, scalable datasets at very low annotation cost.

**Ablations about Image Sources.** In the previous comparisons, the images used in each dataset are not identical. To better control for this variable, we fix the set of images and instead compare the effect of caption quality of different datasets under the Qwen2.5-3B + Qwen2.5-ViT setting. As shown in Table 2, we compare CapRL with ShareGPT4V-1M and DenseFusion-1M. The results demonstrate that, when using the same set of images, further pretraining with the CapRL-3B-annotated dataset enables the LVM to outperform the baselines by more than 2%. This finding indicates that the substantial advantage of the CapRL dataset over the baselines largely stems from the superior quality of its captions, rather than from differences in image diversity.

**Scaling Trend Comparison of Different Datasets.** We further compare the scaling trend of CapRL and DenseFusion under Qwen2.5-3B + Qwen2.5-ViT setting. Specifically, we sample different numbers of image-caption pairs from each dataset for pretraining. As shown in Figure 4, the CapRL dataset consistently outperforms the corresponding DenseFusion dataset across various scales of pretraining data. Moreover, the overall trend indicates that this performance gap continues to widen as the data size increases. This phenomenon highlights the strong scaling properties of the CapRL dataset, thanks to its high-quality captions, LVMs continue to benefit as the dataset size grows.

## 4.2 PRISM SETTING

In the previous section, we demonstrated from the pretraining perspective that captions generated by CapRL are highly beneficial for modality alignment. In this section, we directly evaluate the informativeness of the captions produced by CapRL-3B through the lens of the Decoupled VQA in Prism Framework (Qiao et al., 2024), and compare our CapRL-3B against other captioning models.

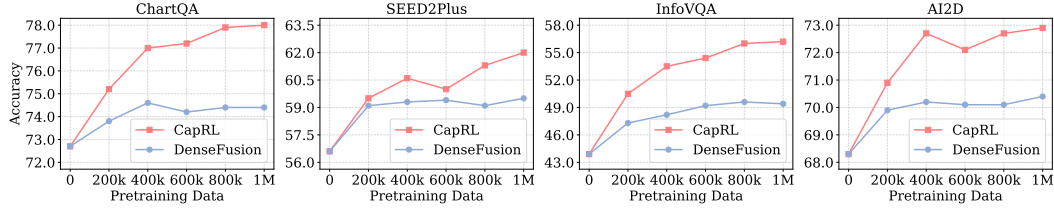


Figure 4: The scaling performance comparison between CapRL-1M and DenseFusion-1M. We use different amounts of pretraining data from the two datasets to observe the scaling trend.

Table 3: Captioning ability comparison in Prism Framework. CapRL-3B achieves comparable performance to Qwen2.5-VL-72B, and significantly surpasses existing strategies that use LVLM-as-a-Judge as the reward. The best results are **bold** and the second-best results are underlined.

Caption Model	GRPO Trained	Chart QA	ChartQA Pro	Info VQA	MMMU Pro	Math Verse	Char Xiv	We Math	Math Vision	MMStar	SEED	MMMU	Average
Qwen2.5-VL-3B	✗	65.6	27.1	40.2	28.6	32.8	21.8	54.4	22.6	46.4	64.1	35.1	39.9
Qwen2.5-VL-7B	✗	74.9	35.4	56.4	30.1	36.4	24.8	57.0	23.3	50.7	67.1	37.9	44.9
Qwen2.5-VL-72B	✗	<u>80.2</u>	<u>38.0</u>	<u>60.8</u>	<b>34.1</b>	<b>39.9</b>	<u>30.7</u>	<b>60.2</b>	<b>24.5</b>	<b>55.0</b>	<u>69.3</u>	<b>39.4</b>	<b>48.3</b>
UnifiedRW-as-Judge-3B	✓	54.9	25.1	33.6	28.1	34.6	20.4	58.2	24.5	45.4	61.2	36.3	38.4
Qwen2.5VL-as-Judge-3B	✓	71.4	34.2	49.3	29.1	33.8	22.9	54.3	24.1	47.7	64.5	36.4	42.5
CapRL-3B	✓	<b>80.5</b>	<b>39.9</b>	<b>64.8</b>	<u>30.7</u>	36.4	<b>32.4</b>	60.1	<u>23.4</u>	<b>55.0</b>	<b>70.6</b>	<u>38.1</u>	<b>48.3</b>

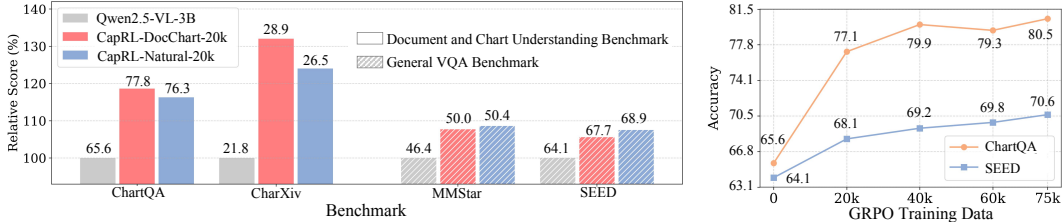


Figure 5: (Left) CapRL demonstrates strong generalization even when trained on images from a single domain. CapRL-DocChart-20k refers to training conducted solely on document or chart images, while CapRL-Natural-20k is trained exclusively on natural images. Both models achieve significant improvements over the baseline on out-of-domain benchmarks, highlighting strong generalization capability. (Right) CapRL demonstrates promising scaling performance on QA training datasets.

**Implementation details.** Similar to our caption reward design, the Prism Framework decouples VQA into two stages. In Stage 1, the captioner generates captions about the input image. In Stage 2, an LLM answers questions based solely on the generated caption. We leverage the Prism framework primarily because it can evaluate caption quality in an objective and stable manner. In our setup, we fix Stage 2 with a fine-tuned Qwen2.5-3B-Instruct as the answering LLM, ensuring that benchmark performance directly reflects the quality of captions produced by the captioner. To assess the effect of different reward designs in GRPO, we include two other baseline models: one trained with UnifiedReward-2.0-qwen-3b (Wang et al., 2025), and the other with Qwen2.5-VL-3B as the judge for caption quality evaluation. The corresponding prompts are provided in Appendix F.

**Comparison with Qwen2.5-VL series.** As shown in Table 3, CapRL-3B significantly outperforms both the 3B and 7B models of the Qwen2.5-VL series, achieving performance comparable to that of the 72B model. In chart and infographic understanding, CapRL-3B surpasses Qwen2.5-VL-3B by 14.9%, 12.8%, and 24.6% on ChartQA, ChartQAPro, and InfoVQA, respectively. For natural image understanding, it leads Qwen2.5-VL-3B by 9.6% and 6.5% on MMStar and SEED. These results demonstrate that GRPO training has substantially unlocked the potential of Qwen2.5-VL-3B, enabling it to fully leverage its inherent knowledge to organize all objects and their attributes within an image into comprehensive and detailed captions. As a result, its perception capability is pushed to the limit, reaching a level comparable to that of the 72B model.

**Comparison with LVLM-as-a-Judge reward.** In our comparison with other reward design methods, we observe that when using UnifiedReward-2.0-qwen-3b as the judge to evaluate caption quality, the model’s captioning ability actually deteriorates during GRPO training. We attribute this to the severe bias present in UnifiedReward-2.0-qwen-3b: during its training, it was exposed to a large number of captions from text-to-image datasets, which are typically short and only describe the main objects. As a result, the UnifiedReward model tends to favor shorter captions. As shown in Fig. 1, the



Table 4: Analysis of the number of QA per image.

Caption Model	ChartQA Pro	Info VQA	MMMU	MMStar	WeMath	Avg
Qwen2.5-VL-3B	27.1	40.2	35.1	46.4	54.4	40.6
CapRL-1QA-20k	35.5	59.8	36.6	50.8	57.3	48.0
CapRL-2QA-20k	36.8	60.2	37.6	51.1	56.6	48.5
CapRL-3QA-20k	36.9	60.3	36.9	51.3	56.8	48.5

Table 5: Ablations about Sampling Rounds N.

Sampling Rounds	ChartQA Pro	Info VQA	MMMU	MMStar	WeMath	Avg
N=1	35.4	58.1	36.5	50.2	56.1	47.3
N=2	36.2	59.1	36.3	49.3	56.9	47.6
N=4	36.7	59.9	37.1	50.9	57.3	48.4
N=8	36.9	59.6	36.5	50.8	57.7	48.3

average caption length during training continuously decreases and eventually collapses to producing only “:description”. Conversely, when using Qwen2.5-VL-3B as the judge, the bias is in the opposite direction: it prefers overly verbose captions. This makes the policy model prone to exploiting the bias by generating long passages of content irrelevant to the image, thereby satisfying the judge model’s preference. As shown in Table 3, the captioning ability under this reward shows significantly inferior to CapRL. Specific examples of such cases are illustrated in Fig. 9, Fig. 10, Fig. 11.

### 4.3 COMPREHENSIVE DISCUSSION ABOUT CAPRL

In this section, we provide a comprehensive analysis and discussion of CapRL. These results further confirm CapRL’s general applicability, robustness, and effectiveness.

**CapRL demonstrates strong generalization even when trained on images from a single domain.** We further investigate the effect of different image sources in the QA dataset used for GRPO training. To this end, we classify the images into two categories using Qwen2.5-VL-3B: (1) documents, charts, or infographics, and (2) natural images. From each category, we sample 20k images for comparison. As illustrated in Fig. 5 (Left), models trained exclusively on chart-type images via GRPO exhibit substantial gains over Qwen2.5-VL-3B, not only in document and chart understanding but also in general VQA tasks. This demonstrates the strong generalization of CapRL-induced captioning improvements beyond the domains encountered during training.

**CapRL demonstrates promising scaling performance on training data.** We conduct training on different amounts of QA data to evaluate the scaling behavior. As shown in Fig. 5 (Right), the model’s performance improves steadily as the amount of QA data increases. These results indicate that our CapRL framework exhibits highly promising scaling potential. With the continued expansion of the training data, the captioning ability can be further enhanced, unlocking additional potential of Qwen2.5-VL-3B. Given its relatively small parameter size and excellent scaling properties, this approach holds strong promise for application in industrial-scale multimodal pretraining.

**Sparse QA supervision is sufficient for CapRL.** We further examine the effect of varying the number of QA pairs per image. Specifically, we randomly sample 20k images that retain three QA pairs after filtering, obtain CapRL-3QA-20k after training. By controlling the number of QA pairs per image, we also construct CapRL-1QA-20k and CapRL-2QA-20k. The results, presented in Table 4, show that even with only a single QA pair per image, Qwen2.5-VL-3B achieves a substantial improvement in captioning performance, averaging 7.4% higher than the baseline and only 0.5% lower than CapRL-2QA-20k. This highlights the remarkable efficiency of CapRL: highly sparse QA supervision is sufficient to unlock significant gains in captioning ability.

**Ablations about sampling rounds N.** Results are shown in Table 5, performance improves steadily when  $N$  increases from 1 to 4, and reaches saturation at  $N = 8$ . The relatively poor performance at  $N = 1$  can be explained by the fact that each question is answered by the LLM only once, without sufficient shuffling of the options. Due to inherent option biases in the LLM, the measured accuracy fails to serve as a reliable proxy for reward, thereby misdirecting the optimization of the policy model.

**Standard Caption Benchmark Evaluation.** We conducted extensive evaluations using ALOHa, CLAIR, and the Factuality and Coverage metrics proposed by CapMAS. Across these metrics, CapRL-3B significantly outperforms Qwen2.5-VL-3B by 7.8% and achieves performance comparable to Qwen2.5-VL-72B. The results in Table 8 highlight that our CapRL algorithm substantially unlocks the potential of Qwen2.5-VL-3B, enabling it to capture image details much more comprehensively.

**Human Study.** To further validate these findings, we randomly sampled 200 images, generating captions using Qwen2.5-VL-3B, Qwen2.5-VL-72B, and CapRL-3B. We invites 9 human judges to evaluate the caption quality based on four criteria: Detail Coverage, Content Accuracy, Expression

Fluency, and Succinctness Degree. As shown in Table 9, CapRL-3B achieves a much higher average score than Qwen2.5-VL-3B and comparable with Qwen2.5-VL-72B.

## 5 CONCLUSION

In this work, we introduce CapRL, a novel framework that successfully applies RLVR to the subjective task of image captioning. By redefining caption quality based on its utility in enabling a vision-free LLM to accurately answer questions, we create a robust, objective reward signal. Our results show that CapRL effectively encourages models to generate dense and precise image descriptions, which in turn substantially promote modality alignment in LVM pretraining. This work marks a significant step away from the restrictive, data-hungry SFT paradigm for RLVR in open-ended tasks.

## ETHICS STATEMENT

Our research does not involve human or animal subjects, and all datasets are either publicly available or collected from the web under open-source licenses. To mitigate potential privacy risks, we applied strict filtering procedures and carried out manual checks to ensure that no sensitive, unsafe, or personally identifiable content was included in the training data. The proposed methodology improves model performance without generating additional manual annotation costs. This reduces reliance on data labeling practices that may involve exploitative labor. We believe the benefits of this work, including enhanced accessibility, more accurate multimodal understanding, and stronger research reproducibility, will bring positive impact to society. For instance, it can help visually impaired individuals better understand a wide range of images. To further encourage responsible deployment, we provide complete documentation of our methods. Our work complies with relevant ethical standards, legal requirements, and the ICLR Code of Ethics.

## REPRODUCIBILITY STATEMENT

We have made significant efforts to ensure the reproducibility of our work. The main paper clearly describes the CapRL framework in Section 3 and the experimental settings in Section 4. Detailed implementation steps, hyperparameters, and training configurations are provided in Appendix J. To support dataset reproducibility, we describe the QA curation process Appendix H, I and the image collection and filtering pipeline Appendix H, which together yield the CapRL-5M dataset. Additional qualitative examples are included in Appendix A to demonstrate model outputs across diverse scenarios. Comprehensive results and ablations are presented in Tables and Figures, enabling transparent comparison with baselines. All proofs of concept are explained with sufficient detail to allow independent replication.

## REFERENCES

- Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S Morcos. Semdedup: Data-efficient learning at web-scale through semantic deduplication. *arXiv preprint arXiv:2303.09540*, 2023.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.
- Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3558–3568, 2021.
- Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for lite vision-language models. *arXiv preprint arXiv:2402.11684*, 2024a.
- Lin Chen and Long Xing. Open-llava-next: An open-source implementation of llava-next series for facilitating the large multi-modal model community. <https://github.com/xiaoachen98/Open-LLaVA-NeXT>, 2024.
- 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 *ECCV*, 2024b.
- Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V Le, Sergey Levine, and Yi Ma. SFT Memorizes, RL Generalizes: A comparative study of foundation model post-training. *arXiv preprint arXiv:2501.17161*, 2025.
- Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving clip training with language rewrites. *Advances in Neural Information Processing Systems*, 36:35544–35575, 2023.

- Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Bing Liu, and Sean Hendryx. Rubrics as rewards: Reinforcement learning beyond verifiable domains. *arXiv preprint arXiv:2507.17746*, 2025.
- 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.
- Alexander Gurung and Mirella Lapata. Learning to reason for long-form story generation. *arXiv preprint arXiv:2503.22828*, 2025.
- Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *CVPR*, 2015.
- Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, et al. Tulu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*, 2024.
- 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.
- Xiaotong Li, Fan Zhang, Haiwen Diao, Yueze Wang, Xinlong Wang, and Ling-Yu Duan. Densefusion-1m: Merging vision experts for comprehensive multimodal perception. *arXiv preprint arXiv:2407.08303*, 2024.
- Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *ACL*, 2004.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NeurIPS*, 2023.
- Zijun Liu, Peiyi Wang, Runxin Xu, Shirong Ma, Chong Ruan, Peng Li, Yang Liu, and Yu Wu. Inference-time scaling for generalist reward modeling. *arXiv preprint arXiv:2504.02495*, 2025a.
- Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. Visual-RFT: Visual reinforcement fine-tuning. In *ICCV*, 2025b.
- Xun Lu. Writing-zero: Bridge the gap between non-verifiable problems and verifiable rewards. *arXiv preprint arXiv:2506.00103*, 2025.
- Run Luo, Lu Wang, Wanwei He, and Xiaobo Xia. Gui-r1: A generalist r1-style vision-language action model for gui agents. *arXiv preprint arXiv:2504.10458*, 2025.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In *ACL*, 2002.
- 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. In *NeurIPS*, 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 *ICML*, 2021.
- Noam Rotstein, David Bensaid, Shaked Brody, Roy Ganz, and Ron Kimmel. FuseCap: Leveraging large language models for enriched fused image captions. In *WCAV*, 2024.
- 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.
- 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.

- Yi Su, Dian Yu, Linfeng Song, Juntao Li, Haitao Mi, Zhaopeng Tu, Min Zhang, and Dong Yu. Crossing the reward bridge: Expanding rl with verifiable rewards across diverse domains. *arXiv preprint arXiv:2503.23829*, 2025.
- Yanpeng Sun, Jing Hao, Ke Zhu, Jiang-Jiang Liu, Yuxiang Zhao, Xiaofan Li, Gang Zhang, Zechao Li, and Jingdong Wang. Descriptive caption enhancement with visual specialists for multimodal perception. *arXiv preprint arXiv:2412.14233*, 2024.
- Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, et al. Kimi K1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*, 2025.
- Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.
- 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 *CVPR*, 2025.
- Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *CVPR*, 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.
- Qiyang Yu, Quan Sun, Xiaosong Zhang, Yufeng Cui, Fan Zhang, Yue Cao, Xinlong Wang, and Jingjing Liu. Capsfusion: Rethinking image-text data at scale. In *CVPR*, 2024.
- Tianyu Yu, Bo Ji, Shouli Wang, Shu Yao, Zefan Wang, Ganqu Cui, Lifan Yuan, Ning Ding, Yuan Yao, Zhiyuan Liu, et al. RLPR: Extrapolating rlvr to general domains without verifiers. *arXiv preprint arXiv:2506.18254*, 2025.

## A CAPRL CASES

We provide further illustrative examples of CapRL-3B to highlight its surprising captioning capabilities in this section.

**Comparison with Qwen2.5-VL-3B.** As illustrated in Fig. 6, CapRL-3B demonstrates remarkable capability in understanding infographics, providing information that is both comprehensive and accurate. In contrast, Qwen2.5-VL-3B, as shown in Fig. 7, makes numerous errors in identifying key information within infographics. Furthermore, Fig. 8 highlights that CapRL-3B achieves substantially higher accuracy in chart understanding compared to Qwen2.5-VL-3B. Similarly, in the case of physical image understanding in Fig. 12, CapRL-3B also demonstrates clear superiority.

**Comparison with UnifiedRW-as-Judge-3B and Qwen2.5VL-as-Judge-3B.** To intuitively illustrate the issues introduced by LVLm-as-a-Judge rewards, we present examples of captions produced by models trained with different methods in Fig. 9, Fig. 11, and 10. The Qwen2.5VL-as-a-Judge-3B model tends to ignore key visual information in the image and outputs lengthy, irrelevant content, such as repeatedly asserting that its caption is of high quality in order to exploit reward hacking. In contrast, UnifiedRW-as-Judge-3B produces overly short captions that omit substantial amounts of critical chart information.

**More cases of CapRL-3B in understanding infographics and natural images.** Fig. 13 and Fig. 14 provide additional evidence of the impressive perceptual capacity demonstrated by CapRL-3B.

## B MORE ANALYSIS EXPERIMENTS ABOUT CAPRL

### Leaking QA Data Leads to Performance Degradation.

We randomly sample 20k instances and construct two training conditions: one using the retained QA and the other using the filtered QA. As shown in Table 6, the model trained on the leaking data performs on average 1.1% worse than the one trained on high-quality data. This indicates that leaking QA introduces spurious reward signals that mislead the optimization of the policy model. Even when captions are not closely aligned with the image content, the LLM may still achieve high answer accuracy, thereby preventing higher rewards from being correctly assigned to genuinely better captions.

Table 6: Comparison between training with data containing leakage issues and training with filtered data. Leaking data leads to an obvious performance drop.

Training Data	ChartQA Pro	Info VQA	MMMU	MMStar	WeMath	Avg
Leaking20k	36.4	58.9	36.1	50.7	55.1	47.4
Refined20k	36.8	60.2	37.6	51.1	56.6	48.5

## C COMPARISON WITH GRPO-TRAINED VQA MODEL

We incorporated a Format Reward during the RL training process on the 20k dataset, ensuring the output adheres to the format: <caption>...</caption><think>...</think><answer>...</answer>. After training with this format, we obtained Reasoning-3B-20k. As shown in the results<sup>7</sup>, while Reasoning-3B-20k achieves a noticeable improvement over Qwen2.5-VL-3B, it falls significantly short of CapRL-3B. Specifically, CapRL-3B surpasses Reasoning-3B-20k by an average of 3.8% across all 11 benchmarks. This substantial margin highlights the superiority of the CapRL algorithm, demonstrating that our method effectively activates the model’s captioning capabilities through sparse QA supervision applied directly to the caption.

Table 7: Comparison with GRPO-trained VQA model.

Caption Model	ChartQA Pro	Info VQA	MMMU	MMStar	WeMath	Average
Qwen2.5-VL-3B	27.1	40.2	35.1	46.4	54.4	40.6
Reasoning-3B-20k	32.7	46.2	36.0	47.2	56.1	43.6
CapRL-3B-20k	36.9	59.6	36.5	50.8	57.7	48.3



**CapRL-3B:**

The image is an infographic titled "Kred 2012: 10 Influential Team USA Olympians," highlighting the most influential Olympic athletes from Team USA in 2012, particularly focusing on their Olympic medals and social media influence scores. Here is a detailed description:

1. **\*\*Header and Awards:\*\***

- The image features three main athletes with their respective medals and Kred scores:
- LeBron James (Gold Medal in Basketball) with a Kred score of 999 and a 6 medal.
- Michael Phelps (Silver Medal in Swimming) with a Kred score of 977 and a 6 medal.
- Hope Solo (Bronze Medal in Soccer) with a Kred score of 977 and a 5 medal.

2. **\*\*Staircase of Medals:**

- The image uses a graphical representation with a blue background adorned with stars, indicative of an American theme. The top shows LeBron James with the highest Kred score, followed by Michael Phelps and Hope Solo with identical scores.

3. **\*\*Social Media Metrics:**

- Followers counts for each athlete:
- Michael Phelps: 689,591 followers
- LeBron James: 5,627,723 followers (highest)
- Hope Solo: 542,813 followers
- Retweeted times and replies received:
- Phelps: 63,983 retweets, 75,695 replies
- James: 1,072,729 retweets, 1,250,351 replies
- Solo: 75,014 retweets, 148,310 replies

4. **\*\*Athletes List with Kred Scores for Different Events:**

- The bottom section lists ten athletes along with their events and Kred scores:

- Durant (Basketball) with 992 and 8 points
- Lochte (Swimming) with 975 and 6 points
- Wieber (Gymnastics) with 937 and 5 points
- Raisman (Gymnastics) with 912 and 6 points
- Douglas (Gymnastics) with 899 and 7 points
- Franklin (Swimming) with 891 and 5 points
- Richards-Roes (Track & Field) with 880 and 8 points

5. **\*\*Celebrity Attention Events:**

- A section highlights which events are winning celebrity attention:
- Wendy Williams (@WendyWilliams) on 30 July: Talking about "GIRL POWER, GO TEAMUSA. #LONDONOLYMPICS" with a link to [usagym.org/pages/post.htm](http://usagym.org/pages/post.htm)
- Samantha Giancola (@MTVsammi) on 30 July: Loving swimming, Soccer track, and gymnastics the best
- Piers Morgan on 31 July: Commenting on Michael Phelps being the "official Greatest medal-winner in Olympic history," pointing to @MichaelPhelps
- Bill Simmons (@sportsguy33) on 30 July: mentioned the Swedish-Norway women's handball match coming up in 30 minutes, noting plenty of seats available on the Swedish bandwagon

6. **\*\*Kred Score Calculation:**

- The image explains that Kred scores are calculated based on influence, which is the ability to inspire action through social media, scored on a 1,000-point scale. Outreach is measured by generosity in engaging with others and helping spread messages through social media.

7. **\*\*Logo and Credits:** The Kr logo is in the bottom left, and the infographic is credited to @Neez duoJindex at the bottom.

Figure 6: An illustrative example of CapRL-3B applied to infographic understanding.

Table 8: Comparison on caption benchmark.

Caption Model	ALOHa	CLAIR	Factuality	Coverage	avg
Qwen2.5-VL-3B	62.2	68.9	79.3	59.3	67.4
Qwen2.5-VL-72B	66.1	80.6	83.8	69.7	75.1
CapRL-3B	65.7	80.7	80.5	73.7	75.2

## D LLM ABLATION IN CAPRL

We employed the Qwen2.5 series as the answer-model, utilizing five different model sizes ranging from 0.5B to 32B to answer questions based on the captions. We observed in 11 that even when using the smallest 0.5B model, CapRL still yields a significant improvement in the captioning capability

**Qwen2.5-VL-3B:**

This image is a graphical representation showcasing the most influential Team USA Olympians in 2012, as ranked by Twitter's "Kred" platform. The ranking is based on both influence and outreach metrics. The image predominantly features athletes from the USA Olympic team, including LeBron James, Hope Solo, Michael Phelps, and others.

At the top of the image, there are three main athletes highlighted with star ratings and scores:

1. **Fred the Teacher**: Has a score of 999 and is rated a "Golden Medal."
2. **Mary Smith**: Has a score of 977 and is rated a "Silent Medal."
3. **Christopher Brown**: Has a score of 977 and is rated a "Subtle Medal."

Below these main athletes, there are:

4. **Frankie Lawrence**: Has a score of 977 and is rated a "Yellow Medal," representing swimming.
5. **Honey CEO**: Has a score of 992 and is rated a "Gold Medal," representing basketball.

Additionally, there is a section titled "Which events and athletes are winning celebrity attention?" where:

- **@ Fisheries**: Talks about the team's performance and the Olympics.
- **@ - fitting Howard**: An article by Howard, discussing the Swedish-Norwegian handball match.
- **@ -Matching**: Discusses the Mexico-US Handball match.

There is also a section explaining "How Kred scores are calculated," which details that influence and outreach are factors in the scoring system, defining it as "the ability to inspire action through social media."

→ **Hallucinations!**

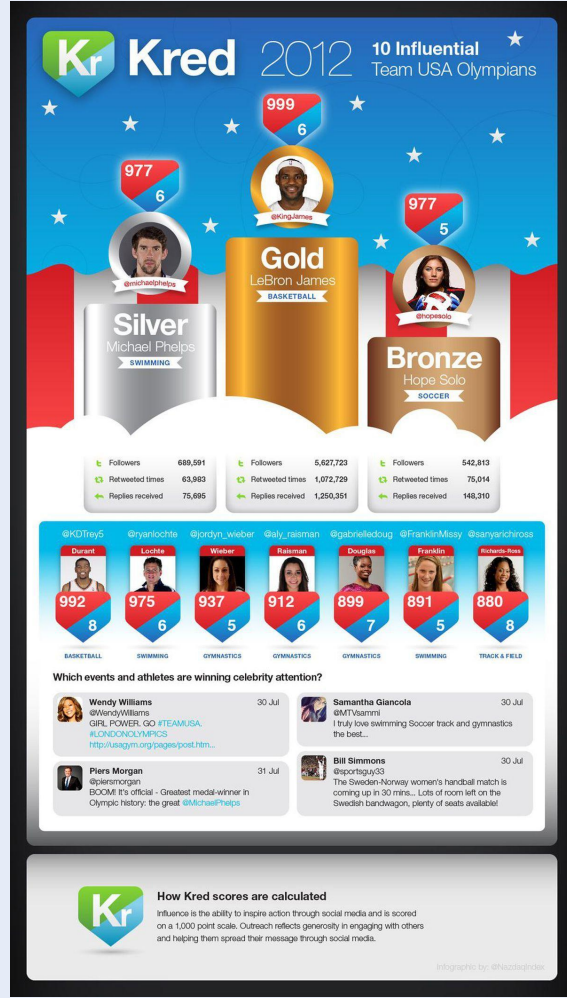


Figure 7: An illustrative example of Qwen2.5-VL-3B applied to infographic understanding.

Table 9: Human study.

Caption Model	Detail Coverage	Content Accuracy	Expression Fluency	Succinctness Degree	Avg
Qwen2.5-VL-3B	3.28	3.45	4.55	3.95	3.81
Qwen2.5-VL-72B	4.32	4.26	4.73	4.21	4.38
CapRL-3B	4.68	4.16	4.65	4.15	4.41

of Qwen2.5-VL-3B; however, there is a noticeable performance drop compared to the 3B model. Performance improves marginally from 1.5B to 3B and then begins to saturate, with no significant variance observed between the 7B and 32B models. These findings indicate that the answer-model does not require excessive parameter size; it simply requires a fundamental capability for information extraction. However, extremely small models, such as the 0.5B variant, appear to lack sufficient capacity for effective extraction.

**CapRL-3B:**

The image is a bar chart that presents survey results regarding opinions on allowing more, about the same, or fewer Syrian refugees into various countries, specifically Jordan, Lebanon, and Turkey. The survey was conducted by the Pew Research Center in Spring 2017 as part of the Global Attitudes Survey, focusing on question Q61. Here is a detailed description:

**\*\*Title:** Jordan, Lebanon, Turkey want fewer Syrian refugees allowed into their country\*\*

**\*\*Question:\*\*** In your opinion, should we allow more refugees from Syria into our country, fewer refugees or about the same as we do now?

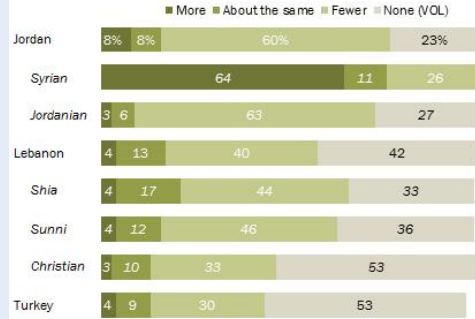
The chart shows percentages for each country and religious or ethnic group regarding their preferred number of refugees:

1. **\*\*Jordan:\*\***
  - 8% want "More"
  - 8% want "About the same"
  - 60% want "Fewer"
  - 23% want "None (VCL)"
2. **\*\*Syrian:\*\***
  - 64% want "More"
  - 11% want "About the same"
  - 26% want "Fewer"
  - (This category includes Syrian respondents overall)
3. **\*\*Jordanian:\*\***
  - 3% want "More"
  - 6% want "About the same"
  - 63% want "Fewer"
  - 27% want "None (VCL)"
4. **\*\*Lebanon:\*\***
  - 4% want "More"
  - 13% want "About the same"
  - 40% want "Fewer"
  - 42% want "None (VCL)"
5. **\*\*Shia:\*\***
  - 4% want "More"
  - 17% want "About the same"
  - 44% want "Fewer"
  - 33% want "None (VCL)"
6. **\*\*Sunni:\*\***
  - 4% want "More"
  - 12% want "About the same"
  - 46% want "Fewer"
  - 36% want "None (VCL)"
7. **\*\*Christian:\*\***
  - 3% want "More"
  - 10% want "About the same"
  - 33% want "Fewer"
  - 53% want "None (VCL)"
8. **\*\*Turkey:\*\***
  - 4% want "More"
  - 9% want "About the same"
  - 30% want "Fewer"
  - 53% want "None (VCL)"

**\*\*Sources:\*\***  
 - Spring 2017 Global Attitudes Survey  
 - Question Q61

**Jordan, Lebanon, Turkey want fewer Syrian refugees allowed into their country**

*In your opinion, should we allow more refugees from Syria into our country, fewer refugees or about the same as we do now?*



Source: Spring 2017 Global Attitudes Survey, Q61.

PEW RESEARCH CENTER

**Qwen2.5-VL-3B:**

...

Here's a summary of the data for each country:

1. **\*\*Jordan\*\*:**
  - More: 8%
  - About the same: 23%
  - Fewer: 68%
2. **\*\*Saudi Arabia\*\*** (assumed to be the "Syrian" bar):
  - More: 64%
  - About the same: 11%
  - Fewer: 25%
3. **\*\*Lebanon\*\*:**
  - More: 40%
  - About the same: 35%
  - Fewer: 25%
4. **\*\*Hijazi Shia\*\*** (assumed as separate):
  - More: 46%
  - About the same: 40%
  - Fewer: 14%
5. **\*\*American Shia\*\*** (assumed as separate):
  - More: 31%
  - About the same: 38%
  - Fewer: 31%
6. **\*\*Sunni Republic\*\*:**
  - More: 20%
  - About the same: 44%
  - Fewer: 36%
7. **\*\*Christian Fiji\*\*:**
  - More: 33%
  - About the same: 38%
  - Fewer: 31%
8. **\*\*Turkey\*\*:**
  - More: 30%
  - About the same: 31%
  - Fewer: 39%

→ Hallucinations!

Figure 8: Chart understanding comparison between CapRL-3B and Qwen2.5-VL-3B.

## E GPT5 EVALUATION

Using the exact same set of images and captions from the Human Study, we employed GPT-5 to score the outputs using a unified prompt. As shown in 10, GPT-5's scoring trends align closely with the

**CapRL-3B:**

The image is a bar chart from the Pew Research Center that illustrates how different groups of Republicans and Republican-leaning individuals view President Trump, specifically focusing on how they see him as "fighting for what I believe in." The chart is broken down by several categories, including age groups, educational levels, political ideology, and whether respondents are full Republicans or leaners towards the Republican Party. Here is a detailed description:

**### Title:**

"Older Republicans especially likely to see Trump as fighting for their beliefs\nAmong Republicans and Republican leaners, % who say the phrase 'fights for what I believe in' describes Trump..."

**### Key Findings:****1. \*\*Overall Percentage (All Rep/Lean Rep):\*\***

- Very well: 51%
- Fairly well: 36%
- NET (Total percentage): 87%

**2. \*\*Age Groups:\*\***

- Ages 18-29: 31% very well, 45% fairly well, total 76%
- 30-49: 41% very well, 42% fairly well, total 82%
- 50-64: 58% very well, 33% fairly well, total 92%
- 65+: 68% very well, 26% fairly well, total 94%

Older age groups (65 and above) are the most likely to see Trump as actively fighting for their beliefs, with the 65+ age group having the highest net percentage at 94%.

**3. \*\*Education Levels:\*\***

- Postgrad: 42% very well, 38% fairly well, total 80%
- College grad: 45% very well, 40% fairly well, total 85%
- Some college: 51% very well, 36% fairly well, total 87%
- HS or less: 56% very well, 33% fairly well, total 89%

Individuals with a high school diploma or less education are most likely to see Trump fighting for their beliefs (89%), while those with postgrad degrees are least likely (80%).

**4. \*\*Political Conservatism:\*\***

- Conserv (Conservative): 63% very well, 31% fairly well, total 94%
- Mod/Lib (Moderate or Liberal): 32% very well, 44% fairly well, total 75%

Conservatives are the most likely to see Trump fighting for beliefs, with a net percentage of 94%.

**5. \*\*Partisanship:\*\***

- Republican: 61% very well, 32% fairly well, total 93%
- Lean Republican: 36% very well, 41% fairly well, total 77%

Leaners towards the Republican Party are less likely to see Trump this way, with a net percentage of 77%.

**6. \*\*Notes:**

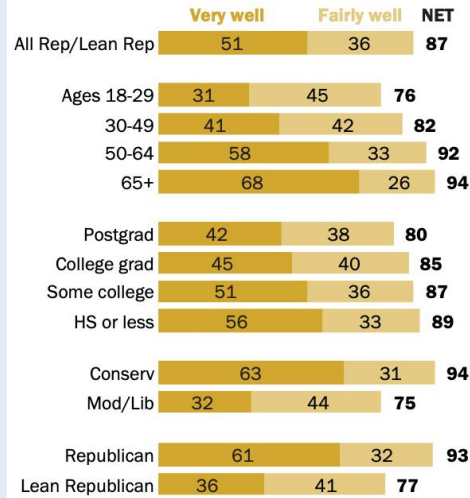
- The survey was conducted February 4-15, 2020, based on Republicans and Republican-leaning independents.
- The Pew Research Center is the source of the data.

**### Percentages Breakdown by Categories:**

- The highest net percentage (94%) is seen in the 65+ age group and Conservatives.
- The lowest net percentage is in the 18-29 age group at 76%.
- Full Republicans (93%) and those leaning Republican (77%) are moderately high.

## Older Republicans especially likely to see Trump as fighting for their beliefs

Among **Republicans and Republican leaners**, % who say the phrase 'fights for what I believe in' describes Trump ...



Note: Based on Republicans and Republican-leaning independents. Source: Survey of U.S. adults conducted Feb. 4-15, 2020.

PEW RESEARCH CENTER

Figure 9: Example of CapRL-3B chart understanding.

human study, once again confirming the strong captioning ability of CapRL-3B. We also observed that GPT-5 appears stricter than human judges regarding Detail Coverage, especially for Qwen2.5-VL-3B, likely because GPT-5 attends to more fine-grained visual details than human evaluators. The prompt is provided in Figure 19.



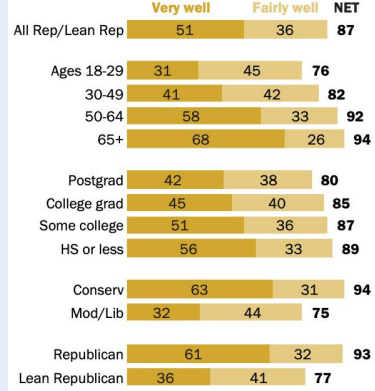
**UnifiedRewardModel-as-Judge-3B:**

This image appears to be a bar chart from a study conducted by the Pew Research Center. The title "Older Republicans especially likely to see Trump as fighting for their beliefs" is prominently displayed at the top. The chart compares percentages among Republicans and Republican leaners on how well the phrase 'fights for what I believe in' describes Donald Trump, across different categories. Key data points are listed with "Very well," "Fairly well," and "NET" columns for each category. Specific numbers are provided for age groups (e.g., All Rep/Lean Rep: 51 for 'Very well', 36 for 'Fairly well'), education levels (postgrad, college grad, etc.), political orientation (conservative, moderate/liberal), and Republican affiliation (Republican, Lean Republican). There are numerical values and percentages (e.g., 87, 92, 85) indicating net results for each category. The source and survey details (U.S. adults, Feb. 4-15, 2020) are noted at the bottom. The description captures the structured, comparative data format and the specifics of the study's findings in detail.

**Image-related key information is incomplete and inaccurate.**

**Older Republicans especially likely to see Trump as fighting for their beliefs**

Among **Republicans and Republican leaners**, % who say the phrase 'fights for what I believe in' describes Trump ...



Note: Based on Republicans and Republican-leaning independents. Source: Survey of U.S. adults conducted Feb. 4-15, 2020.

PEW RESEARCH CENTER

Figure 10: Example of Qwen2.5VL-as-Judge-3B chart understanding.

Table 10: GPT5 Evaluation.

Caption Model	Detail Coverage	Content Accuracy	Expression Fluency	Succinctness Degree	Avg
Qwen2.5-VL-3B	2.85	3.08	4.30	4.02	3.56
Qwen2.5-VL-72B	4.21	4.42	4.59	4.12	4.34
CapRL-3B	4.38	4.39	4.49	4.05	4.33

## F PROMPT USED

We provide all the prompts employed in our experiments in this section. Specifically, the prompt used in CapRL for guiding the LLM to answer questions conditioned on captions is illustrated in Figure 15; the prompt for utilizing the Unified Reward Model as the reward model is shown in Figure 16; and the prompt for adopting Qwen2.5-VL-3B as the reward model is presented in Figure 17.

## G LARGE LANGUAGE MODEL USAGE

In the process of preparing this paper, we made use of large language models (LLMs) primarily as a writing assistant rather than as a methodological component. Specifically, LLMs were employed to refine the readability and fluency of the manuscript by polishing grammar, improving phrasing, and enhancing stylistic consistency across sections. This usage allowed us to present the technical content with greater clarity and precision, ensuring that the intended contributions could be more effectively communicated to the reader. Moreover, the models were used to check for potential linguistic errors or awkward constructions that might distract from the scientific substance of the work. Importantly, we emphasize that the role of LLMs in this study was strictly limited to text editing and proofreading. They were not involved in designing the research methodology, formulating experimental procedures, analyzing results, or drawing conclusions. All conceptual development, data processing, modeling, and evaluation were carried out independently by the authors. By constraining the role of LLMs to language refinement, we aimed to benefit from their strengths in natural language processing while maintaining full intellectual control over the scientific and technical aspects of this research.

Table 11: LLM ablation in CapRL.

Caption Model	ChartQA Pro	InfoVQA	MMMU	MMStar	WeMath	Average
Qwen2.5-VL-3B	27.1	40.2	35.1	46.4	54.4	40.6
Answer0.5b	33.7	54.9	35.8	48.6	56.1	45.8
Answer1.5b	36.2	59.5	36.7	50.7	56.6	47.9
Answer3b	36.9	59.6	36.5	50.8	57.7	48.3
Answer7b	35.7	60.2	36.9	51.0	56.9	48.1
Answer32b	35.9	60.1	35.7	50.2	56.3	47.6

## H DATA PROCESSING

To ensure both quality and safety, our data processing pipeline consists of three main stages. First, inspired by SemDeDup (Abbas et al., 2023), we construct clusters to identify and remove images with redundant semantics. During this step, we also discard low-resolution and overly simple images, while filtering out content that involves violence, pornography, or other safety concerns. Second, to avoid benchmark leakage, we integrate the images used in commonly referenced evaluation datasets and form clusters with them. Any images from our collection that are overly similar to benchmark samples are eliminated. Third, we conduct a safety inspection through human verification. Annotators perform sample-based screening, and once the proportion of unsafe images falls to a negligible level, we stop filtering. Following this process, we obtain the final dataset, CapRL-5M.

## I QA PROCESSING

In constructing the QA pairs, we employ the Qwen2.5-VL-72B model with prompts shown in Figure 18. For each image, we generate five questions and retain those without leakage issues. We do not deliberately control the number of QA pairs per image, prioritizing instead the overall dataset size and diversity. As revealed in later ablation studies 4, although even a single QA per image proves highly effective, adding more QA pairs still brings marginal improvements.

During QA filtering, since the model’s answers carry uncertainty due to temperature parameter, we do not filter questions solely based on the correctness of a single response. Instead, we sample responses four times for each question, shuffling the answer options each time, and then measure the accuracy of the LLM’s answers both based on the image and based only on the question itself. We ultimately apply a threshold to filter out questions with high image-based accuracy but low question-only accuracy.

It is worth noting that, because our filtering criteria are quite strict, some discarded QA pairs contain only mild or even negligible leakage. This also explains why, as shown in Figure 6, training with the leaked data does not cause training collapse but merely led to degraded performance.

## J PRETRAINING DETAILS

**Model Architecture.** In our experimental setup, the language model component is initialized with a pretrained LLM, the visual encoder is initialized with a pretrained ViT, and the MLP projector is randomly initialized. This setup corresponds to a commonly adopted starting point in multimodal pretraining. To ensure the robustness of our conclusions, we evaluate three groups of architectures: (1) Qwen2.5-3B + Qwen2.5-ViT, (2) Qwen2.5-7B + Qwen2.5-ViT, and (3) InternLM2.5-7B + CLIP-ViT-L. This selection jointly considers differences in parameter scale, LLM backbone, and visual encoder type.

**Training Setting.** Following the training paradigm of ShareGPT4V, our training process consists of three stages: Initial Alignment – Further Pretraining – SFT. (1) In the Initial Alignment stage, we unfreeze the MLP and perform preliminary alignment using the BLIP-558K dataset. We adopt a learning rate of 1e-3 and a batch size of 256. (2) In the Further Pretraining stage, we unfreeze all parameters including the LLM, MLP, and ViT. This stage facilitates further alignment with various



high-quality image-caption datasets, enabling the LLM to better understand visual features. We set the learning rate to  $4e-5$  and the batch size to 256. (3) In the SFT stage, we again unfreeze all parameters and train on the OpenLLaVA-Next dataset. We set the learning rate to  $2e-5$  and the batch size to 128.

**Baselines.** We selected several strong baselines for comparison. (1) Vanilla, which skips the Further Pretraining stage and only goes through the first and third stages. Additionally, we constructed two more baselines by varying the dataset used in the Further Pretraining stage: (2) ShareGPT4V-1M, and (3) DenseFusion-1M. To ensure a fair comparison by controlling the number of samples, we randomly sampled 1 million image-caption pairs from the 5M dataset to form CapRL-1M.

**Qwen2.5VL-as-Judge-3B:**

This image is a detailed bar graph from a survey conducted by the Pew Research Center, which examines how Republicans and Republican-leaning individuals view Donald Trump's political leadership using a specific phrase, "fights for what I believe in." Here's a comprehensive description and analysis:

1. **Title and Subject**: The study asks how Republicans and Republican-leaning individuals view Donald Trump, focusing on how they see him "fighting" for their beliefs.

2. **Survey Data**: The data is divided into several categories and metric variables, including political ideologies, and par

**Image-related key information is incomplete and inaccurate.**

3. **Key Observations**:

- Among All Rep/Lean Reps, 87% see Trump as very well or fairly well described as someone who "fights for what I believe in."
- Younger age groups (18-29 and 30-49) have lower percentages (76% for ages 18-29 and 82% for ages 30-49) compared to the older groups (94% for ages 50-64 and 92% for ages 65+).
- Education level: Postgrads see Trump this way 80% of the time, while those with some college education see him 87% of the time.
- Political ideology: Conservative individuals see Trump this way 94% of the time, with moderate/liberal (75% of the time).

4. **Age and Hubri**:

- Younger politically-aware generations (e.g., 18-29, 30-49) show less association with Trump's leadership, indicating a generational shift in perception.
- Older generations (50-64 and 65+) associate Trump more strongly with this phrase, suggesting a deeper understanding of his motives.

5. **Educational Background and Political Affiliation**:

- Higher levels of education correlate with a stronger perception of Trump's commitment to his beliefs.
- Conservative individuals overwhelmingly see Trump as fighting for their beliefs, which contradicts or signifies a significant portion of their loyalty.

**Image-unrelated information is verbose and lengthy**

6. **Party Affiliation**:

- Republicans and Republican-leaning individuals see Trump as fighting for their beliefs 93% of the time, with leaning Republicans (36% very well, 41% fairly well) joining this perception essentially.

7. **Percentages and Interpretation**:

- The highest percentage (94%) seeing Trump in this light belongs to conservative respondents, indicating their strong belief in his ideological commitment.
- The lowest percentage (36%) is among moderate/liberal Republicans, signaling that these individuals feel less connected to Trump's political ethos compared to the conservative and leaning Republican groups.

8. **Survey Context**:

- The survey was conducted in February 2020, reflecting current political sentiment and personal belief perceptions about Trump's leadership style.

9. **Possible Implications**:

- This indicates changes in age-related political denominations within the Republican party, with older generations having a more consistent and steadfast belief in Trump's dedication to his principles.
- Education appears to serve as a crucial factor in shaping how individuals perceive Trump's political commitment.
- There is a strong correlation between party affiliation and belief in Trump's ideological fervor.

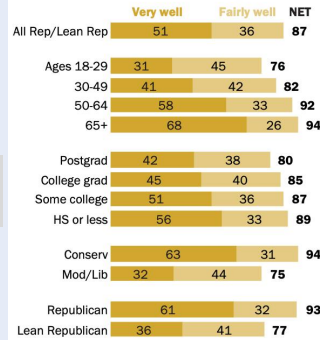
10. **Conclusion**: The data highlights the diverse perceptions of how Donald Trump represents his targets' beliefs, showing variations by age, education, and political alignment. It also underscores the strong ideological and short-term political connection among conservatives, while revealing generational and educational shifts impacting views on Trump's leadership style.

This detailed description includes all aspects of the image, providing a comprehensive understanding of the data and its implications. It covers various key demographic categories, their perceptions, and the overall narrative they present, enabling a pure text model to answer a wide range of questions related to this graph effectively. The explanation is structured logically, covering different dimensions of the data, from the percentages observed to the implications of these observations, thereby offering a thorough analysis of the given information. The description also includes relevant inferences and conclusions, which are derived from the data presented. The use of text blocks and a structured format aids clarity and ease of comprehension, making it easy for a text model to understand and effectively generate answers to queries based on this description. The inclusion of societal and political context enriches the understanding of the image, providing a broader perspective on the subject matter. This comprehensive approach ensures that the description is both informative and engaging, covering all necessary details and providing a well-rounded understanding of the graph.

This level of detail also enables other AI systems to process and analyze the information accurately, ensuring that they can answer questions such as: What age groups see Trump most strongly? How does educational background influence perception? What is the political affinity link? These questions can be answered based on the coherent and detailed description provided. The description is based on all visible information in the image, ensuring that no part of the content is overlooked or misunderstood. The structured and logical flow makes it easy to follow, with each segment contributing to a comprehensive understanding of the graph and its implications. The description is both factual and insightful, providing a robust basis for any further inquiries or question-answering-related tasks. Thus, it serves as an excellent example of how to approach and describe an image with precision and thoroughness.

**Older Republicans especially likely to see Trump as fighting for their beliefs**

Among Republicans and Republican leaners, % who say the phrase 'fights for what I believe in' describes Trump ...



Note: Based on Republicans and Republican-leaning independents. Source: Survey of U.S. adults conducted Feb. 4-15, 2020.

PEW RESEARCH CENTER

Figure 11: Example of Qwen2.5VL-as-Judge-3B chart understanding.

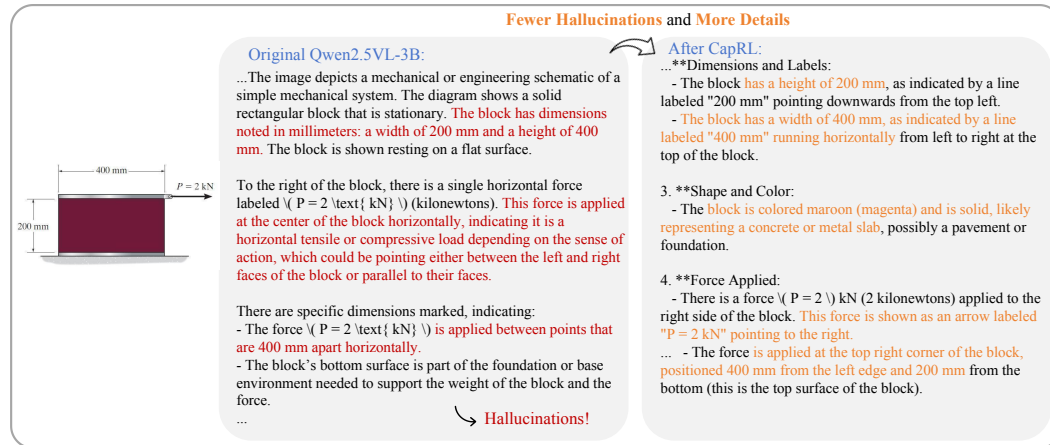


Figure 12: Example of CapRL enhancing the captioning ability of Qwen2.5-VL-3B.

**CapRL-3B:**

The image is an infographic titled "Social Media Statistics: Breaking Down the Major Social Networks," specifically focusing on Facebook, Twitter, and Pinterest. Below is a detailed description:

---

**\*\*Social Media Statistics\*\***

**\*\*BREAKING DOWN THE MAJOR SOCIAL NETWORKS\*\***

### ### Facebook

- **\*\*77%\*\***: B2C companies have acquired customers through Facebook.
- **\*\*43%\*\***: B2B companies have acquired customers through Facebook.
- Facebook accounts for 1/2 of all retargeted clicks on the web.
- 189 million "mobile only" users and 488 million mobile users.
- 23% login 5 or more times per day.
- 47% of users say Facebook influences purchases.
- 80% of social media users prefer connecting with brands on Facebook.
- Facebook is made up of 43% male and 57% female users.

### ### Twitter

- 55-64 largest growing age demographic, up 76% since 2013.
- 500 million daily users, producing 175 million tweets per day.
- 69% of all reported follows are suggested by friends.
- 34% of web marketers have generated leads using Twitter.
- 32% of all web users are using Twitter.
- 56% of all customer tweets to brands are ignored. Research shows the highest CTR (Click-through Rate) comes between 4:30 and 5:30 PM EST.
- A green icon with 56% highlights this statistic.

### ### Pinterest

- 80% of registered users are female (specifically highlighted with a icon).
- 80% of all pins are "re-pins."
- Pinterest referrals spend 70% more than non-social traffic.
- Pins with listed prices get 36% more likes than those without.
- With 17MIL+ engagements, only 15% are on brand-specific boards.
- 69% have purchased or want to purchase via Pinterest.

### ### Additional Details about Pinterest:

- Key points about Pinterest:
  - 80% of all pins are "re-pins."
  - Pinterest referrals are highly effective.
  - Pins with prices perform better, getting 36% more likes.
  - Only 15% of engagements are on brand-specific boards.
  - 69% of users have purchased or want to purchase via Pinterest.

### ### Footer Information:

- Queen City Media specializes in brand and content marketing on all major social media platforms. Contact them today for a FREE social media evaluation.
- Contact details:
  - [www.qcmny.com](http://www.qcmny.com)
  - [info@qcmny.com](mailto:info@qcmny.com)
  - Phone: 716-845-6500
  - Address: 515 Main St, Buffalo, NY 14203
- Social media links provided:
  - [fb.com/queencitymedia](https://fb.com/queencitymedia)
  - [twitter.com/qcmny](https://twitter.com/qcmny)
  - [pinterest.com/qcmny](https://pinterest.com/qcmny)

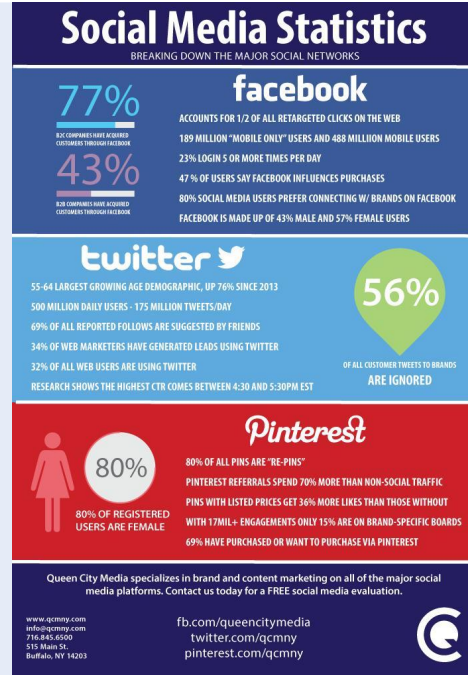


Figure 13: An illustrative example of CapRL applied to infographic understanding.

**CapRL-3B:**

The image depicts a serene outdoor scene featuring two individuals, likely a couple, sitting on a wooden bench in what appears to be a garden or a courtyard decorated in patriotic themes, probably for a historical or commemorative event given the context. Here is a detailed description:

**1. \*\*Setting and Location:\*\***

- The image is taken in a garden or patio area, evidenced by the presence of plants, flowers, and a garden bench. The setting is adorned with British flags, specifically Union Jacks, suggesting a British context, possibly near a location in the United Kingdom.

**2. \*\*Subjects:\*\***

- The image shows a man and a woman sitting side by side on a wooden bench.

- The woman is on the left, wearing a green dress that is a vintage style, likely from the 1940s or 1950s, given its emerald green color and full skirt. She has blonde hair and is wearing sunglasses and dark brown shoes. She is holding the man's hand.

**3. \*\*Attire:\*\***

- The woman is dressed in a green dress, which is greenish-blue in color, typical of a retro or wartime era outfit, possibly from the 1940s.

**4. \*\*Man:\*\***

- The man is on the right, dressed in a military uniform, specifically a cap (possibly a sailor's cap) and a dark blue military uniform with anchor pins, indicating he might be a naval or air force officer. He is wearing a cap with a badge and a jacket with buttons, typical of military attire from that period.

**5. \*\*Furniture and Background:\*\***

- They are sitting on a wooden bench that is outdoors, likely in a garden or park, as indicated by the various potted plants and flowers around it. The bench is outdoors, possibly on a sidewalk or patio.

**6. \*\*Flag:\*\***

- A large British flag (Union Jack) is prominently displayed behind the couple, draped over the back of the bench. The flag is red, white, and blue, with remnants of text or inscriptions on it, suggesting it might be old or vintage, possibly from a ship or military context. This flag spans across the image from left to right.

**7. \*\*Flag Display: Union Jack:\*\***

- The flag is large and appears to be old, with hues of red, white, and blue, typical of the British national flag. It has a mix of pink in the middle, indicating it might be from a historical or naval setting.

**8. \*\*Plants and Decor:\*\***

- The garden is lush with various plants and flowers, including a pot on the left containing greenery in a blue pot, likely herbs or small plants. There are also other planted areas with tall green onions or garlic chives (likely *Allium fistulosum*).

**9. \*\*Ground and Additional Elements:\*\***

- The ground is paved with a brick or concrete sidewalk, and there is a small drain or cover on the right side of the image. The bottom left corner mentions "Mercury Press & Media Ltd," suggesting the image is from a media source.

**10. \*\*Text and Settings:\*\***

- The sky is bright, indicating it is a sunny day, likely spring or summer given the well-maintained garden and the couple's vintage clothing.

**11. \*\*Objects:\*\***

- The man is wearing dark navy blue pants and shoes typical of military uniformity, complete with belt and decorations typical of a naval officer.



Figure 14: An illustrative example of CapRL applied to natural image understanding.

### Prompt for LLM to answer questions based on Caption

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 End>

Question: {}

Figure 15: Prompt for LLM to answer questions based on Caption.

### Prompt for Unified Reward Model as a Judge

You are presented with an image along with its generated text caption. Your task is to comparatively evaluate the caption quality based on the image.

Provide overall assessment for the caption (rated from 1 to 5):

- Overall Score: How well the caption comprehensively covers the information in the image, provides detailed descriptions of each object, and maintains accuracy in the conveyed information.

Output your evaluation using the format below:

- Overall Score (1-5): X

Your task is provided as follows:

Text Caption: [{}]

Figure 16: Prompt for Unified Reward Model as a Judge.



### Prompt for Qwen2.5-VL-3B as a Judge

You are presented with an image along with its generated text caption. Your task is to comparatively evaluate the caption quality based on the image.

Provide overall assessment for the caption (rated from 1 to 5):

- Overall Score: How well the caption comprehensively covers the information in the image, provides detailed descriptions of each object, and maintains accuracy in the conveyed information.

Output your overall score for this caption as a single number.

Your task is provided as follows:

Text Caption: [{}]

Your evaluation:

Figure 17: Prompt for Qwen2.5-VL-3B as a Judge.

### Prompt Qwen2.5-VL-72B to generate QA

Your task is to generate five multiple-choice questions and their answers about the object based on the provided image.

The questions should be challenge and focus on the image content, you answer should strictly follow the following format:

#### 1. \*\*Which method achieves the highest accuracy (Acc) on the FF++ (HQ) dataset?\*\*

- A) Method "a"
- B) Method "b"
- C) Method "c"
- D) Ours

**\*\*Answer:\*\* D) Ours**

-----

#### 2. \*\*What is the primary color of the kayak in the image?\*\*

- A) Blue
- B) Red
- C) Black
- D) White

**\*\*Answer:\*\* B) Red**

-----

You should strictly follow the above format and should not generate irrelevant sentences. All the question should be answered based on the image.

Figure 18: Prompt Qwen2.5-VL-72B to generate QA

1458  
1459  
1460  
1461  
1462  
1463  
1464  
1465  
1466  
1467  
1468  
1469  
1470  
1471  
1472  
1473  
1474  
1475  
1476  
1477  
1478  
1479  
1480  
1481  
1482  
1483  
1484  
1485  
1486  
1487  
1488  
1489  
1490  
1491  
1492  
1493  
1494  
1495  
1496  
1497  
1498  
1499  
1500  
1501  
1502  
1503  
1504  
1505  
1506  
1507  
1508  
1509  
1510  
1511

## Prompt for GPT5 Evaluation

You are a specialist in multimodal evaluation. Your responsibility is to judge the quality of a candidate image–text description strictly based on the visible content of the provided image. All assessments must rely solely on what is visually verifiable—no external knowledge, assumptions, or contextual guessing is permitted.

You will score the description across **four analytical dimensions**, each on a **1–5 scale**, where higher scores indicate stronger performance.

Every score must be supported with clear and concise reasoning.

---

### **## Evaluation Dimensions**

#### **### 1. Detail Coverage**

Assesses whether the description captures the key visual elements present in the image.

Check whether major objects, attributes, spatial relations, and contextual cues are appropriately observed and represented.

#### **### 2. Content Accuracy**

Evaluates the factual correctness of the description with respect to the image.

Verify that all mentioned entities, colors, quantities, positions, and visual facts align with what is actually visible, without introducing contradictions.

#### **### 3. Expression Fluency**

Measures the clarity, coherence, and naturalness of the language.

Determine whether the description is well-structured, grammatically sound, and easy to read.

#### **### 4. Succinctness Degree**

Judges whether the description conveys the necessary information in a concise and efficient manner.

Assess whether the text avoids unnecessary verbosity while still preserving completeness and informativeness.

### **## Scoring Scale (1–5)**

#### **### 1 — Critical flaws**

Major inconsistencies, severe inaccuracies, missing essential content, or highly speculative interpretation.

#### **### 2 — Significant deficiencies**

Multiple noticeable issues; partial alignment with the image but clearly below acceptable quality.

#### **### 3 — Adequate**

Meets basic expectations; generally correct but contains clear room for refinement or more precise articulation.

#### **### 4 — Strong**

Mostly accurate, well-structured, and coherent; only minor issues or slight omissions.

#### **### 5 — Excellent**

Comprehensive, precise, fluent, fully aligned with the visual evidence, and free of unjustified inference.

---

### **## Output Format (JSON)**

The output format must strictly follow the form below. Do not output anything else.

```
{
  "detail_coverage": ,
  "content_accuracy": ,
  "expression_fluency": ,
  "succinctness_degree": ,
}
```

Figure 19: Prompt GPT5 to evaluate caption