BalCapRL🧢: A Balanced Framework for RL-Based MLLM Image Captioning

Shaokai Ye Vasileios Saveris Yihao Qian Jiaming Hu Elmira Amirloo
Peter Grasch
Apple

Abstract

Image captioning is one of the most fundamental tasks in computer vision. Owing to its open-ended nature, it has received significant attention in the era of multimodal large language models (MLLMs). In pursuit of ever more detailed and accurate captions, recent work has increasingly turned to reinforcement learning (RL). However, existing captioning-RL methods and evaluation metrics often emphasize a narrow notion of caption quality, inducing trade-offs across core dimensions of captioning. For example, utility-oriented objectives can encourage noisy, hallucinated, or overlong captions that improve downstream question answering while harming fluency, whereas arena-style objectives can favor fluent but generic descriptions with limited usefulness. To address this, we propose a more balanced RL framework that jointly optimizes utility-aware correctness, reference coverage, and linguistic quality. In order to effectively optimize the resulting continuous multi-objective reward formulation, we apply GDPO-style reward-decoupled normalization to continuous-valued captioning rewards and show that it improves performance over vanilla GRPO. Additionally, we introduce length-conditional reward masking, yielding a more suitable length penalty for captioning. Across LLaVA-1.5-7B and Qwen2.5-VL 3B and 7B base models, our method consistently improves caption quality, with peak gains of +13.6 DCScore, +9.0 CaptionQA, and +29.0 CapArena across different models.

1 Introduction

Image captioning is a fundamental visual task. Early captioning models tend to generate short descriptions centered around closed-vocabulary objects. Advances in MLLMs [Liu et al., 2024, Qwen Team et al., 2025] have enabled increasing open-ended and detailed captions.

In order to maximize captioning capabilities of modern MLLMs, reinforcement learning aimed at improving captioning performance as an objective (captioning-RL) [Xing et al., 2025, Huang et al., 2026, Ye et al., 2025] has increasingly gained popularity.

Existing captioning-RL methods often optimize a narrow notion of caption quality and then evaluate improvements using benchmarks aligned with that same perspective. We find that this creates a systematic bias: gains on one dimension of caption quality often come with regressions or moderate improvements on others.

We identify three major views that currently shape captioning-RL and caption evaluation: downstream utility [Yang et al., 2025], correctness-and-completeness with respect to reference captions [Ye et al., 2025], and arena-style preference judgments [Cheng et al., 2025]. Each captures an important aspect of caption quality, but optimizing any one view in isolation is insufficient. Correctness-and-coverage objectives can reward repetitive, mechanical and rigid descriptions. Utility-oriented training can

40th Conference on Neural Information Processing Systems (NeurIPS 2026).



Figure 1: **Top.** Illustration of captions from different biased models. **Bottom.** Different captioning-RL models evaluated on benchmarks representing different views. Δ (higher is better) denotes the difference between the RL-trained model and the base QwenVL2.5-3B model. CapRL and RubiCap results are produced by evaluating official checkpoints.

encourage hallucinated or overly long captions that help downstream question answering while degrading fluency. Arena-style judgments, in contrast, can favor fluent yet generic captions that rank well in CapArena despite being less useful and less informative. As a result, prior methods often exhibit clear trade-offs across benchmarks rather than uniformly better captioning. Figure 1 illustrates this pattern for representative prior methods, as well as purposefully biased variants of our method, obtained by removing individual components from our framework. To address this issue, we propose BalCapRL, a more balanced reinforcement learning framework for detailed image captioning. Our method jointly optimizes rewards for utility-aware correctness, reference-coverage completeness, and linguistic quality. Because these reward dimensions can have distinct and partially competing optimization dynamics, we find that vanilla GRPO is suboptimal in our setting, and therefore apply GDPO [Liu et al., 2026] to continuous-valued rewards, which we refer to as c-GDPO, to better optimize multi-reward policy optimization (Figure 2). Additionally, we introduced a novel two-sided length penalty via reward masking, which we show is better suited to captioning-RL.

Across LLaVA-1.5-7B [Liu et al., 2024] and QwenVL2.5 3B and 7B [Qwen Team et al., 2025], BalCapRL consistently improves caption quality across benchmarks representing all three views, outperforming prior methods in almost all settings. These results suggest that better captioning-RL requires not just optimization toward a single benchmark, but a more balanced training objective that explicitly accounts for multiple dimensions of caption quality.

In summary, our contributions are as follows:

- We identify and empirically characterize a systematic bias in captioning-RL: optimizing a single view (utility, correctness-and-completeness, or arena) regresses the others (Figure 1).
- We propose a balanced multi-dimensional reward consisting of pointability-aware precision, reference-coverage recall, and linguistic quality, computed via a single MLLM judge without separately trained reward models.
- We introduce the pointability rubric, a novel criterion that rejects true-but-non-pointable meta-commentary and closes a reward-hacking loophole of the standard precision reward.
- We extend GDPO's reward-decoupled normalization to continuous, competing multi-rewards (c-GDPO), with a supporting analysis (Proposition 1).

- We introduce two-sided length-conditional reward masking, which is suited to captioning by preventing both over-length captions and length collapse.
- We propose b-CapScore, a balanced single-number captioning metric with high human alignment.

2 Related work

Reinforcement learning for MLLMs on image captioning. Reinforcement learning has been used in MLLMs to mitigate hallucinations [Sun et al., 2023], improve multimodal reasoning [Feng et al., 2025], and better align model outputs with human preferences [Yu et al., 2025a]. More generally, reinforcement learning with verifiable rewards (RLVR) [Shao et al., 2024] is effective when the target outcome can be checked automatically. For tasks without simple verifiable outcomes, prior works often rely on LLM/MLLM feedback signals [Lee et al., 2024, Gunjal et al., 2025, Yu et al., 2025a]. Recently, captioning-centric RL work [Ye et al., 2025, Xing et al., 2025, Huang et al., 2026, Zhang et al., 2025, Tang et al., 2026] has gained increasing attention. FEEDQUILL [Ye et al., 2025] decomposes generated captions into atomic assertions and uses MLLMs to assess whether each assertion is visually verifiable or covered by the reference captions. It then trains two separate reward models from preference data and uses them as RL reward signals. Similar to FEEDQUILL, our method also incorporates both precision and recall into the reward. However, unlike FEEDQUILL, which requires two separate reward models, our approach directly uses MLLM-judge which greatly simplifies the pipeline while introducing a latency and cost tradeoff. In addition, we include a pointability rubric to the precision reward and include linguistic reward to discourage repetitive content that degrades fluency and coherence. CapRL [Xing et al., 2025] uses a two-stage pipeline in which generated captions are sent to a text-only LLM for question answering, and answer accuracy is used as the reward. In contrast, we avoid explicit MQA sampling and instead encourage caption utility through a simple notion of pointability. This provides finer-grained supervision at the caption level and reduces the risk of over-optimizing for downstream MQA performance alone. RubiCap [Huang et al., 2026] uses a committee of strong MLLMs to construct per-sample rubrics for rubric-based reinforcement learning. In contrast, our method does not require multiple MLLMs to generate sample-specific rubrics. We also explicitly incorporate a utility-oriented reward component, which leads to improved caption utility compared to RubiCap.

Image Captioning metrics. Traditional image captioning metrics, including BLEU [Papineni et al., 2002], METEOR [Banerjee and Lavie, 2005], and CIDEr [Vedantam et al., 2015], measure similarity between generated captions and human references using n-gram overlap. While widely used, such metrics are sensitive to surface-level phrasing variations and therefore struggle to capture the diversity of valid descriptions. Model-based metrics, such as SPICE [Anderson et al., 2016], CAPTURE [Pothiraj et al., 2025], and CLIPScore [Hessel et al., 2022], were introduced to alleviate this limitation. However, recent studies show that these metrics become less reliable for the long, detailed, and nuanced captions produced by modern MLLMs in open-ended captioning settings. To evaluate MLLM captions, recent work increasingly relies on LLMs or strong commercial MLLMs as judges. These approaches can be broadly categorized into three perspectives: (1) **the utility view**, which measures how well a caption supports downstream text-only question answering, as in Prism [Qiao et al., 2024] and CaptionQA [Yang et al., 2025]; (2) **the correctness-and-completeness view**, which evaluates faithfulness and coverage of the captions [Ye et al., 2025, Jing et al., 2024, Liu et al., 2025a]; and (3) **the arena view**, which assesses caption quality through pairwise competition, as in CapArena [Cheng et al., 2025]. Our work argues that these aspects should be combined rather than considered separately, thereby reducing the biases introduced by any single evaluation aspect.

3 Method

3.1 Reward design

To obtain scalar reward signals for caption quality, we first incorporate the correctness-and-completeness perspective. Our method is related in spirit to FEEDQUILL [Ye et al., 2025] in that both decompose captions into atomic assertions for rewards derived from precision and recall. Specifically, we decompose both the policy generated caption as well as a ground-truth caption into atomic assertions to enable the computation of precision and recall, which provide reward

signals for correctness and completeness, respectively. Unlike FEEDQUILL, our method does not require training separate reward models; instead, we compute the rewards directly via judge-based decomposition and verification, yielding a simpler and more modular pipeline. However, as discussed in Section 1, correctness and completeness alone are insufficient: they do not prevent captions from being correct yet not useful, nor do they prevent degradation in fluency. We therefore introduce two additional components: a pointability principle, used as a rubric to constrain what is considered as a useful atomic assertion, and a linguistic score that regularizes the model toward fluent and coherent captions.

Decomposition. Given a model-generated caption C, we employ a large language model (LLM) to decompose it into a set of **atomic assertions**:

$$\mathcal{A} = \{a_1, a_2, \ldots, a_N\}, \tag{1}$$

where N denotes the total number of atomic assertions extracted from the caption. Similarly, the reference caption (data generation details in Appendix A.1) is decomposed into a set of reference units:

$$\mathcal{O} = \{o_1, o_2, \ldots, o_M\}, \tag{2}$$

where M is the number of reference units.

Precision (Utility-aware Correctness). The precision reward R_{prec} measures the proportion of atomic assertions in $\mathcal{A}$ that are **verifiably correct**. An atomic assertion $a_i \in \mathcal{A}$ is considered a **true positive** if and only if it satisfies the following two conditions:

1. **Visually verifiable** ($\mathcal{A}_G$): The assertion can be verified as factually correct from the image content by a vision-language model (VLM).
2. **Pointability** ($\mathcal{A}_P$): The assertion refers to a visually pointable element—specifically, something that a person can physically point to in the image. Compared with prior work [Ye et al., 2025], this novel addition specifically discourages generating non-pointable, low utility meta commentary. An empirical example is shown in Figure 3 and the prompt is provided in Appendix A.2.

Formally, let $\mathcal{A}^+ \subseteq \mathcal{A}$ denote the set of true positive assertions:

$$\mathcal{A}^+ = \mathcal{A}_G \cap \mathcal{A}_P, \tag{3}$$

where $\mathcal{A}_G = \{a_i \in \mathcal{A} \mid a_i \text{ is visually verified}\}$ and $\mathcal{A}_P = \{a_i \in \mathcal{A} \mid a_i \text{ is pointable}\}$.

The precision reward is then computed as: $R_{\text{prec}} = \frac{|\mathcal{A}^+|}{|\mathcal{A}|}$.

Recall (Reference Coverage). The recall reward R_{rec} measures the extent to which the model-generated caption covers the key information present in the reference caption. We employ an LLM to perform the matching, assessing whether each atomic assertion $o_j \in \mathcal{O}$ is mentioned or can be reasonably inferred from the generated atomic assertions (Appendix A.2).

Let $\mathcal{Q} = \mathcal{A} \cap \mathcal{O}$ denote the set of matched units between the generated and reference captions, as determined by the LLM. The recall reward is computed as $R_{\text{rec}} = \frac{|\mathcal{Q}|}{|\mathcal{O}|}$.

Linguistic Score. The linguistic reward R_{ling} evaluates the linguistic quality of generated captions using an LLM (Appendix A.2) that assesses three dimensions: *Clarity*, measuring readability and absence of ambiguity; *Fluency*, evaluating grammatical correctness and natural phrasing; and *Coherency*, assessing logical flow and unified structure. Each of the three dimensions is normalized to the range [0,1], and the final linguistic reward is computed as their average.

3.2 Data

Across experiments, we use images from ShareGPT4V [Chen et al., 2023]. The dataset contains roughly 90K image-text pairs, originally captioned by GPT-4V [OpenAI et al., 2024]. We re-captioned the data with GPT-5-mini [Singh et al., 2025], reusing the original captioning prompts, and use these updated reference captions for our main results.

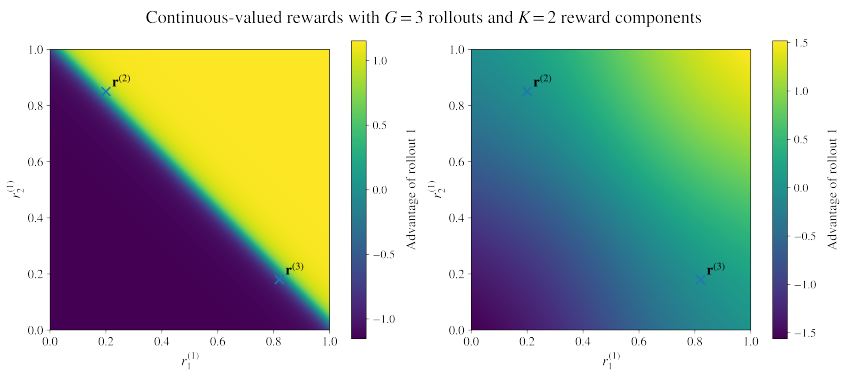

Figure 2: **Illustration of summed-reward collapse under vanilla GRPO.** Left: when rewards are aggregated before group normalization, the normalized advantage depends only on the aggregated reward, so reward vectors with identical aggregates are indistinguishable, and nearby aggregates may be only weakly separated. Right: c-GDPO normalizes rewards separately before aggregation, avoiding this invariance and yielding more distinguishable advantages in the same setting. We vary rollout 1 over $[0, 1]^2$ while fixing $r^{(2)} = (0.20, 0.85)$ and $r^{(3)} = (0.82, 0.18)$.

3.3 Policy Optimization

Applying GDPO to Continuous Captioning Rewards. Recently, GRPO [Shao et al., 2024] and its variants [Liu et al., 2026, Yu et al., 2025b, Gao et al., 2025, Liu et al., 2025b] have become widely used policy optimization methods. In the original GRPO and most of its follow-ups, when multiple rewards are present, these rewards are first summed and then group-normalized, which can lead to the collapse of distinct rollout advantages [Liu et al., 2026]. GDPO [Liu et al., 2026] addresses this issue by decoupling normalization across reward dimensions.

We observe that this pathology is not limited to discrete or verifiable reward settings. In our setting, the precision, recall, and linguistic rewards are continuous-valued with distinct dynamics. Nevertheless, vanilla GRPO still sums these rewards before group normalization, reducing each rollout to a single scalar. As a result, the resulting advantage depends only on a one-dimensional projection of the reward vector, causing distinct continuous reward trade-offs to become indistinguishable when their aggregated rewards coincide. Thus:

Proposition 1. *Consider a K-reward, G-rollout setting with continuous-valued rewards, where $G \geq 3$ and each reward dimension has nonzero within-group variance. Under vanilla GRPO, if rewards are aggregated before group normalization, then for any fixed competing rollouts, the normalized advantage of a rollout depends on its reward vector only through its aggregated reward. Consequently, all reward vectors lying on the same aggregated-reward hyperplane are indistinguishable to the optimizer. In contrast, reward-decoupled normalization computes per-reward normalized deviations before aggregation, and therefore is not invariant to these hyperplanes.*

Proof is provided in Appendix A.3.

Therefore, following the same decoupled-normalization principle as GDPO, we apply it to continuous-valued multi-reward optimization, which we refer to this continuous-reward instantiation as c-GDPO. We do not claim novelty in the normalization operator itself: c-GDPO is a direct application and theoretical extension (Proposition 1) of GDPO to the continuous, competing multi-reward setting. This enables its application to our precision, recall, and linguistic rewards, while preserving finer distinctions among different reward combinations and providing more expressive training signals.

To illustrate the effect of c-GDPO in the continuous-valued setting, Figure 2 shows that vanilla GRPO loses fine-grained multi-reward signal after reward aggregation and normalization. In particular, in the saturated region, different underlying reward combinations can yield nearly identical advantage values. In contrast, c-GDPO preserves these differences in the final advantage (Figure 2), which leads to more stable optimization in our setting.

Specifically, given a group of G rollouts for each input prompt, we first compute the normalized advantage for each reward. For the j-th rollout, the individual advantages are:

$$A^{(j)}_{\text{prec}} = \frac{R^{(j)}_{\text{prec}} - \mu_{\text{prec}}}{\sigma_{\text{prec}} + \epsilon}, \quad A^{(j)}_{\text{rec}} = \frac{R^{(j)}_{\text{rec}} - \mu_{\text{rec}}}{\sigma_{\text{rec}} + \epsilon}, \quad A^{(j)}_{\text{ling}} = \frac{R^{(j)}_{\text{ling}} - \mu_{\text{ling}}}{\sigma_{\text{ling}} + \epsilon}, \tag{4}$$

where $R^{(j)}_{\text{prec}}$, $R^{(j)}_{\text{rec}}$, and $R^{(j)}_{\text{ling}}$ are the precision, recall, and linguistic rewards of the j-th rollout; μ_{prec}, μ_{rec}, μ_{ling} and σ_{prec}, σ_{rec}, σ_{ling} denote the mean and standard deviation of each respective reward across all G rollouts in the group; and ϵ is a small positive constant for numerical stability (e.g., when all rollouts in a group receive the same reward). For brevity, we omit the prompt index i in Eqs. (4)–(6); when it is needed, as in the objective below, we write $\hat{A}^{(i,j)}_{\text{sum}}$ for the advantage of the j-th rollout of the i-th prompt.

The overall advantage is then obtained by a weighted sum of the normalized advantages:

$$A^{(j)}_{\text{sum}} = w_{\text{prec}} A^{(j)}_{\text{prec}} + w_{\text{rec}} A^{(j)}_{\text{rec}} + w_{\text{ling}} A^{(j)}_{\text{ling}}, \tag{5}$$

where w_{prec}, w_{rec}, and w_{ling} are hyperparameters controlling the relative importance of each reward objective.

Finally, a batch-level normalization is applied:

$$\hat{A}^{(j)}_{\text{sum}} = \frac{A^{(j)}_{\text{sum}} - \mu_{\text{batch}}}{\sigma_{\text{batch}} + \epsilon}, \tag{6}$$

where μ_{batch} and σ_{batch} are computed over all rollouts in the training batch.

Let D denote the captioning training dataset. The corresponding multi-reward GDPO objective is:

$$\mathcal{J}_{\text{GDPO}}(\theta) = \mathbb{E}\left[\frac{1}{G}\sum_{j=1}^{G}\sum_{t=1}^{|o_{i,j}|} \min\Big(s_{i,j,t}\hat{A}^{(i,j)}_{\text{sum}}, \text{clip}(s_{i,j,t}, 1-\epsilon, 1+\epsilon)\hat{A}^{(i,j)}_{\text{sum}}\Big)\right], \tag{7}$$

where $\mathbb{E}$ is over $(q_i, o_i) \sim D$ and $o_{i,j} \sim \pi_{\theta_{\text{old}}}(\cdot \mid q_i)$, and $s_{i,j,t} = \pi_\theta(o_{i,j,t} \mid q_i, o_{i,j,<t}) / \pi_{\theta_{\text{old}}}(o_{i,j,t} \mid q_i, o_{i,j,<t})$ is the token-level importance sampling ratio. We provide more details in Appendix A.1.

Length-Conditional Reward Masking. Recently there has been increasing work to train reasoning models to add length constraints [Liu et al., 2025c, Team et al., 2026] for token efficiency. However, length constraints in training models with captioning-based RL serve different purposes: models could produce excessively long captions, possibly containing redundant information, in an effort to increase recall, or conversely aim to maximize precision by reducing caption length, thus missing key information. Therefore the length constraint in captioning objectives cannot be one-sided (upper bound only) as commonly done in reasoning models.

In the presence of a reference caption from either a human or a strong reference captioning model, a natural choice for a length constraint is to constrain the generated caption length with respect to the reference caption. For example, one may use the ratio between the generated caption and the reference caption as a linear length penalty. However, such a linear length penalty can limit the exploration by prematurely encouraging the model to converge its generation length to the reference caption length, which is especially amplified when the reference caption has a very different length compared to the policy model's original caption length.

To avoid restricting exploration in the early stages of training, we instead introduce length-conditional reward masking that acts as a gating mechanism. Let ℓ_{pred} and ℓ_{ref} denote the token lengths of the predicted and reference captions, and define the length ratio as $\rho = \ell_{\text{pred}}/\ell_{\text{ref}}$. Let τ_l and τ_u ($0 \le \tau_l < \tau_u$) denote the lower and upper thresholds on this length ratio, respectively. The linguistic reward is then masked by $\tilde{R}_{\text{ling}} = \begin{cases} R_{\text{ling}}, & \text{if } \tau_l \le \rho \le \tau_u, \\ 0, & \text{otherwise.} \end{cases}$

4 Results

4.1 Experiments

As discussed in Section 1, considering only one aspect of captioning risks introducing bias. We use DCScore [Ye et al., 2025] to represent the correctness-and-completeness view, CaptionQA [Yang

Table 1: Main results.

Model	DCScore ↑		CaptionQA ↑		CapArena ↑		Arena Length	b-CapScore ↑	
Proprietary MLLMs									
GPT-4o	45.8	–	78.7	–	10.3	–	83	56.1	–
Gemini-3.1-Flash	59.9	–	83.9	–	81.2	–	197	65.6	–
GPT-5.4	66.5	–	89.2	–	82.2	–	220	70.9	–
Gemini-3.1-Pro	67.8	–	90.0	–	79.8	–	362	73.0	–
LLaVA-1.5-7B									
Baseline	23.0	–	46.4	–	-94.0	–	74	26.9	–
FEEDQUILL	34.5	+11.5	-	-	-	-	-	–	–
Ours	**36.6**	+13.6	**55.4**	+9.0	**-65.0**	+29.0	124	**43.4**	+16.5
QwenVL2.5-3B									
Baseline	43.3	–	70.0	–	-34.0	–	131	46.2	–
CapRL-3B	48.6	+5.3	**82.6**	+12.6	-50.6	-16.6	403	47.4	+1.2
RubiCap-3B	43.0	-0.3	71.1	+1.1	-29.5	+4.5	149	46.8	+0.6
Ours	**50.8**	+7.5	75.0	+5.0	**-3.8**	+30.2	175	**49.3**	+3.1
QwenVL2.5-7B									
Baseline	46.0	–	74.9	–	13.7	–	136	51.1	–
RubiCap-7B	50.5	+4.5	76.0	+1.1	22.7	+9.0	176	53.9	+2.8
Ours	**53.4**	+7.4	**79.1**	+4.2	**28.5**	+14.8	192	**58.7**	+7.6

Table 2: Performance comparison across vision benchmarks.

Model	BLINK	ChartQA	DocVQA	InfoVQA	MMBench	MMStar	OCRBench	ScienceQA	SEEDBench	TextVQA	Avg.
QwenVL2.5-3B	47.91	83.28	**92.85**	**74.22**	55.88	55.79	78.70	83.35	74.94	**78.62**	72.55
ShareGPT5V-mini-SFT	45.11	82.60	85.92	69.94	53.92	56.19	79.80	**83.47**	74.51	66.87	69.83
RubiCap-3B	45.38	82.84	90.15	71.26	51.96	56.29	75.90	82.60	**75.77**	75.13	70.73
CapRL-3B	47.88	83.40	90.18	73.17	**57.84**	**56.94**	**81.00**	83.40	75.67	73.26	72.27
BalCapRL-3B	**47.92**	**83.64**	92.78	74.17	56.86	55.93	78.70	83.31	75.55	78.40	**72.73**

et al., 2025] to represent the utility view and CapArena [Cheng et al., 2025] to represent the arena view. We also report average caption length on CapArena as an additional indicator of model behavior. Additionally, we introduce b-CapScore, a balanced captioning metric that takes harmonic mean of pointability-aware precision, reference coverage, and linguistic quality; its definition and human-alignment analysis are in Appendix A.5.

In the main results, we test our method with LLaVA1.5-7B, and QwenVL2.5 series of 3B and 7B model sizes. We compare our model with the base QwenVL2.5 models and three recent captioning-RL methods, FEEDQUILL [Ye et al., 2025], CapRL [Xing et al., 2025] and RubiCap [Huang et al., 2026]. We report FEEDQUILL numbers for LLaVA1.5-7B from their paper as the checkpoint is not released. CapRL is evaluated at 3B size as only 3B size is available. For RubiCap, we evaluate both the 3B and 7B checkpoints.

Next, we perform leave-one-out ablations to assess the contribution of each component, through which we identify the causes of specific biased behaviors. Additionally, we include ablation studies to investigate the impact of reward weight (Appendix A.6), impact of using different training-time MLLM judges (Appendix A.7) and additional qualitative examples (Appendix A.8).

4.2 Main results

BalCapRL consistently outperforms prior work in captioning benchmarks. As shown in Table 1, with LLaVA-1.5-7B, our method strongly improves over the baseline across all metrics, lifting DCScore by 13.6 points, CaptionQA by 9.0 points, and CapArena by 29.0 points. Compared to FEEDQUILL, which arguably optimizes DCScore, our method still outperforms it by 2.1 points on this metric.

Using QwenVL2.5-3B as the base model, compared to CapRL-3B, our method strongly outperforms it in DCScore and CapArena, though CapRL-3B scores higher still in CaptionQA. Notably, CapRL-

Table 3: Standard captioning performance on the COCO caption benchmark (higher is better).

Model	BLEU-4 ↑	CIDEr ↑	METEOR ↑	ROUGE-L ↑
QwenVL2.5-3B	**31.4**	**104.0**	28.1	**53.3**
CapRL-3B	24.5	84.8	**28.5**	51.6
RubiCap-3B	23.9	81.3	28.4	50.9
BalCapRL-3B	29.8	99.3	28.2	52.6

3B produces captions roughly 3× longer than the base policy, and even regresses compared to it in CapArena by 16.6 points. We believe this to be a direct result of CapRL's optimization method, which strictly optimizes captions for MQA utility, leading to excessively long captions with degraded fluency as also demonstrated in the qualitative examples in Appendix A.8. In comparison, our method's balanced objective improves over the baseline across all metrics.

Compared to RubiCap-3B, our method strongly outperforms it on all metrics, even approaching the larger RubiCap-7B in DCSCore and CaptionQA performance. When using the same QwenVL2.5-7B base model, our method again significantly outperforms RubiCap-7B on all evaluated benchmarks.

BalCapRL largely preserves general vision benchmark performance. Beyond captioning benchmarks, we also study the performance of models on ten general vision benchmarks in Table 2. We first create a baseline that finetunes the base model using Supervised Fine-tuning (SFT) on the same RL training data (referred as ShareGPT5V-mini) and then evaluate captioning-RL models. The results show that while SFT improves the performance in some benchmarks such as ScienceQA, it loses nontrivial performance in most benchmarks compared to its base model, confirming the findings of RubiCap [Huang et al., 2026]. Surprisingly, while RL is commonly believed to suffer less from catastrophic forgetting, prior work such as RubiCap and CapRL models still suffer from the regression, potentially due to their imbalanced reward design. We note that CapRL-3B significantly improves performance in MMBench and OCRBench while suffering from nontrivial degradation in TextVQA and DocVQA. In contrast, BalCapRL has no notable regression in any tested benchmark while having improvements on several benchmarks.

BalCapRL better preserves short-caption capability. While the current trend of captioning is to move to more detailed ones, keeping model's capability for shorter captions is still very relevant to many downstream applications. We show in Table 3 the traditional metrics on COCO caption. All listed finetuned QwenVL2.5-3B via captioning-RL has a regression on short-caption performance. However, BalCapRL better preserves the short-captioning capability compared to CapRL and RubiCap.

The proposed method is robust to the tested judge choices. We show in Appendix A.7 the method remains effective when varying the choice of the judge models (i.e., GPT-4o-mini, GPT-5-mini, GPT-5.4). Note that results in Table 1 were obtained via GPT-4o-mini judge for the low cost and fast training, while stronger judge such as GPT-5.4 could yield even better results.

4.3 Ablation Studies

Table 4: Leave-one-out ablation studies. w_{prec}, w_{rec}, w_{ling} are set equal for this study.

Model	DCScore	CaptionQA	CapArena	Arena Length
QwenVL2.5-3B	43.3	70.0	-34.0	131
Full	52.0	75.0	-12.0	203
w/o c-GDPO	38.0	67.0	-71.8	96
w/o Precision (CapArena-biased)	41.2	73.6	-13.8	163
w/o Recall	51.8	74.9	-19.3	152
w/o Linguistic (Utility-biased)	53.4	76.3	-51.0	375
w/o Pointability	39.2	63.5	-85.7	240
w/o recap with gpt-5-mini	46.8	73.6	-17.3	152

To assess the impact of each component of our method, we start with leave-one-out ablation studies in Table 4, followed by an ablation study on length penalty in Table 5. We focus on QwenVL2.5-3B as it is the most commonly used model among prior captioning-RL work.

Table 5: Ablation studies over choice of length penalty and hyperparameters.

Model	DCScore	CaptionQA	CapArena	Arena Length
QwenVL2.5-3B	43.3	70.0	-34.0	131
w/o length penalty	39.0	71.6	-32.5	81
w/ linear length penalty	52.7	72.7	-33.0	90
w/ proposed length penalty	52.0	75.3	-19.0	203
Length-Conditional Reward Masking (τ_l, τ_u)				
$\tau_l = 0, \tau_u = 3$	39.4	71.4	-31.0	80
$\tau_l = 0.5, \tau_u = 1$	46.1	74.0	-17.7	121
$\tau_l = 0.5, \tau_u = 2$	52.0	75.3	-19.0	203
$\tau_l = 0.5, \tau_u = 3$	54.0	75.9	-62.7	313
$\tau_l = 0.5, \tau_u = 4$	55.4	75.9	-54.0	322
$\tau_l = 0.5, \tau_u = 5$	54.8	75.5	-72.0	355
$\tau_l = 0.5, \tau_u = 6$	54.5	75.8	-36.7	280

Keeping c-GDPO is critical in our setting. We first examine the effect of removing c-GDPO. Instead of applying separate group normalization to each reward, as in c-GDPO, we follow vanilla GRPO by summing the three rewards and applying group normalization only to the aggregated reward. Relative to the QwenVL2.5-3B baseline, vanilla GRPO leads to substantial performance degradation across benchmarks. We attribute this to the vanilla GRPO's difficulty in learning fine-grained signals of multiple rewards with distinct dynamics (Figure 2).

Effects of removing individual rewards. We then perform an ablation study by removing each of the three rewards from the full method (Table 4), setting the corresponding reward weight to zero. Removing the precision reward maintains the large gain in CapArena but leads to a clear drop in DCScore. We denote this variant as the CapArena-biased model in Figure 1. This behavior is expected: without the precision constraint, the model can more freely optimize toward matching the reference captions (generated by GPT-5-mini) and improving linguistic score, even when doing so reduces visually-verifiable precision and harms DCScore. In contrast, removing the recall reward yields performance above the baseline on all benchmarks and remains only slightly below the full method. This suggests that our framework still improves the model without relying on the recall reward. Notably, removing the linguistic reward increases both CaptionQA and DCScore compared to even our full method, but causes a substantial drop in CapArena. We label this variant, which somewhat resembles CapRL in behavior, as the utility-biased model in Figure 1. Similar to CapRL-3B, we observe an approximately $3\times$ increase in caption length, suggesting that the model may generate repetitive or overly enumerative content at the expense of fluency and coherence. These results highlight that linguistic quality is not adequately captured by CaptionQA and DCScore alone, and should therefore be explicitly considered in captioning RL.

Keeping the pointability rubric mitigates meta commentary. Removing the pointability rubric largely hurts the model (Table 4). A qualitative examination in Figure 3 shows that without pointability rubric, the model learns to hack the rewards by abusing the use of meta-commentary (fluent but lacking downstream utility).

Better reference caption helps. We examine the impact of reference-caption quality by replacing the GPT-5-mini recapped captions with the original ShareGPT4V captions (see Table 4). This results in consistent drops across benchmarks, suggesting that higher-quality reference captions provide a more effective recall signal under our framework. Notably, this variant performs slightly worse than removing the recall reward altogether. This suggests that a weak or poorly aligned reference-caption signal may be less beneficial than omitting the recall objective entirely, under our framework.

Length penalty ablation studies. We study the impact of the length penalty in Table 5. We begin by removing the length penalty from our framework. Under this setting, the model falls into a suboptimal regime in which it regresses in DCScore and slightly improves in CaptionQA and CapArena. Note that its caption length becomes shorter than its base model, potentially exploiting shorter captions to avoid hallucinations too aggressively. This is undesirable as we expect to have the model generate more detailed captions with the same or fewer hallucinations. We then add a length penalty that penalizes deviations of the predicted-to-reference caption length ratio from a predefined acceptable interval in a piecewise linear manner (Appendix A.9). This greatly improves DCScore, suggesting the importance of length penalty for captioning-RL. Then we show that our length-conditional reward masking outperforms the linear length penalty in a more balanced way and resulting in longer caption

length. Next we study our length-conditional reward masking by varying τ_l and τ_u. We observed that setting an upper bound τ_u is not sufficient to prevent the model from generating short captions to over-avoid mistakes. In contrast, setting τ_l to 0.5 effectively mitigates this behavior. Next we fix τ_l to 0.5 and study the effect of varying τ_u from 1 to 6. We observe that allowing the model to generate longer captions would generally improve CaptionQA and DCScore, but at the expense of CapArena performance.

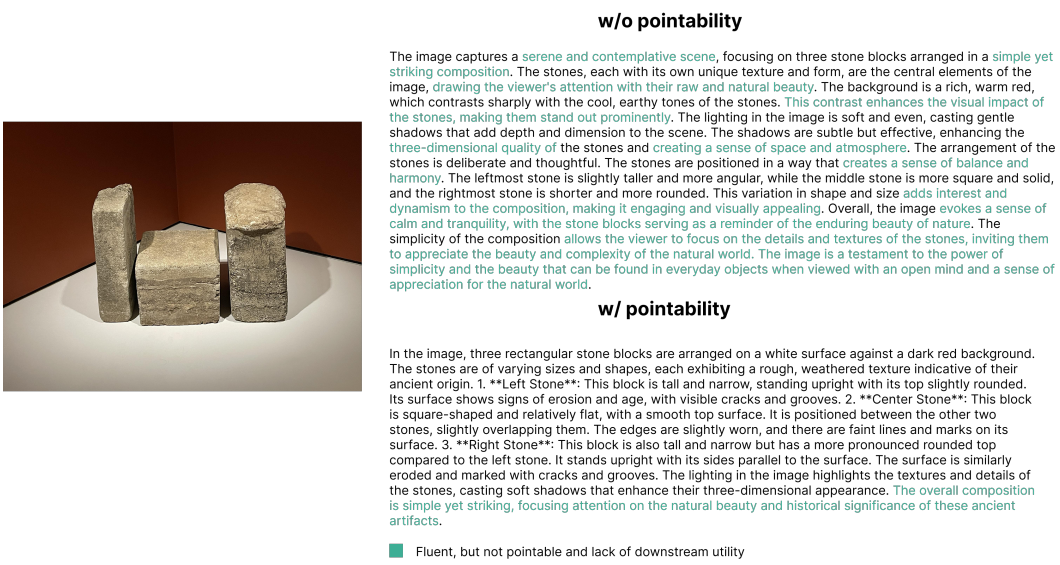

Figure 3: **A qualitative comparison between with pointability rubric and without.**

5 Limitations

One limitation of our work is that some aspects of caption quality are not explicitly modeled in our reward design. In particular, plausible inferences supported by world knowledge may be under-rewarded by the pointability rubric, which favors visually pointable and directly verifiable content. Such information can only be indirectly preserved through the recall objective, making performance on these aspects dependent on the quality and coverage of the reference captions. As a result, our framework may undervalue captions that contain reasonable but non-pointable inferences. Another limitation is our method uses MLLM-as-judge, which largely simplifies the pipeline compared to related method FEEDQUILL but introduces latency and cost tradeoff.

6 Conclusion

We showed that existing captioning-RL methods, by optimizing a single view of caption quality, systematically trade one dimension for another: utility-oriented objectives yield overly long and less fluent captions, correctness-and-completeness objectives yield mechanical descriptions, and arena-style objectives favor fluent but generic captions. To address this, we proposed BalCapRL, a balanced RL framework that jointly optimizes pointability-aware precision, reference-coverage recall, and linguistic quality using a single MLLM judge. We further applied reward-decoupled normalization to continuous, competing rewards (c-GDPO) and introduced two-sided length-conditional reward masking. Across LLaVA-1.5-7B and QwenVL2.5 3B and 7B, BalCapRL consistently improves caption quality on benchmarks representing all three views, while largely preserving general vision and short-caption capabilities. We hope these findings encourage future captioning-RL work to explicitly account for multiple dimensions of caption quality rather than optimizing toward a single benchmark.

References

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2024. URL https://arxiv.org/abs/2310.03744.

Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 Technical Report, 2025. URL https://arxiv.org/abs/2412.15115.

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, 2025. URL https://arxiv.org/abs/2509.22647.

Tzu-Heng Huang, Sirajul Salekin, Javier Movellan, Frederic Sala, and Manjot Bilkhu. RubiCap: Rubric-guided reinforcement learning for dense image captioning, 2026. URL https://arxiv.org/abs/2603.09160.

Qinghao Ye, Xianhan Zeng, Fu Li, Chunyuan Li, and Haoqi Fan. Painting with words: Elevating detailed image captioning with benchmark and alignment learning, 2025. URL https://arxiv.org/abs/2503.07906.

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?, 2025. URL https://arxiv.org/abs/2511.21025.

Kanzhi Cheng, Wenpo Song, Jiaxin Fan, Zheng Ma, Qiushi Sun, Fangzhi Xu, Chenyang Yan, Nuo Chen, Jianbing Zhang, and Jiajun Chen. CapArena: Benchmarking and analyzing detailed image captioning in the llm era, 2025. URL https://arxiv.org/abs/2503.12329.

Shih-Yang Liu, Xin Dong, Ximing Lu, Shizhe Diao, Peter Belcak, Mingjie Liu, Min-Hung Chen, Hongxu Yin, Yu-Chiang Frank Wang, Kwang-Ting Cheng, Yejin Choi, Jan Kautz, and Pavlo Molchanov. GDPO: Group reward-decoupled normalization policy optimization for multi-reward rl optimization, 2026. URL https://arxiv.org/abs/2601.05242.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning large multimodal models with factually augmented rlhf, 2023. URL https://arxiv.org/abs/2309.14525.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in MLLMs, 2025. URL https://arxiv.org/abs/2503.21776.

Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai Dang, Taiwen He, Xiaocheng Feng, Jun Song, Bo Zheng, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Rlaif-v: Open-source ai feedback leads to super gpt-4v trustworthiness, 2025a. URL https://arxiv.org/abs/2405.17220.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models, 2024. URL https://arxiv.org/abs/2402.03300.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. RLAIF vs. RLHF: Scaling reinforcement learning from human feedback with ai feedback, 2024. URL https://arxiv.org/abs/2309.00267.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean Hendryx. Rubrics as rewards: Reinforcement learning beyond verifiable domains, 2025. URL https://arxiv.org/abs/2507.17746.

Lin Zhang, Xianfang Zeng, Kangcong Li, Gang Yu, and Tao Chen. Sc-captioner: Improving image captioning with self-correction by reinforcement learning, 2025. URL https://arxiv.org/abs/2508.06125.

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, 2026. URL https://arxiv.org/abs/2602.21655.

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 on Association for Computational Linguistics*, ACL '02, page 311–318, USA, 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://doi.org/10.3115/1073083.1073135.

Satanjeev Banerjee and Alon Lavie. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Jade Goldstein, Alon Lavie, Chin-Yew Lin, and Clare Voss, editors, *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics. URL https://aclanthology.org/W05-0909/.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation, 2015. URL https://arxiv.org/abs/1411.5726.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: Semantic propositional image caption evaluation, 2016. URL https://arxiv.org/abs/1607.08822.

Atin Pothiraj, Elias Stengel-Eskin, Jaemin Cho, and Mohit Bansal. Capture: Evaluating spatial reasoning in vision language models via occluded object counting, 2025. URL https://arxiv.org/abs/2504.15485.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning, 2022. URL https://arxiv.org/abs/2104.08718.

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, 2024. URL https://arxiv.org/abs/2406.14544.

Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. FaithScore: Fine-grained evaluations of hallucinations in large vision-language models, 2024. URL https://arxiv.org/abs/2311.01477.

Zhihang Liu, Chen-Wei Xie, Bin Wen, Feiwu Yu, Jixuan Chen, Pandeng Li, Boqiang Zhang, Nianzu Yang, Yinglu Li, Zuan Gao, Yun Zheng, and Hongtao Xie. Capability: A comprehensive visual caption benchmark for evaluating both correctness and thoroughness, 2025a. URL https://arxiv.org/abs/2502.14914.

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, 2023. URL https://arxiv.org/abs/2311.12793.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel

Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL `https://arxiv.org/abs/2303.08774`.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, Alex Makelov, Alex Neitz, Alex Wei, Alexandra Barr, Alexandre Kirchmeyer, Alexey Ivanov, Alexi Christakis, Alistair Gillespie, Allison Tam, Ally Bennett, Alvin Wan, Alyssa Huang, Amy McDonald Sandjideh, Amy Yang, Ananya Kumar, Andre Saraiva, Andrea Vallone, Andrei Gheorghe, Andres Garcia Garcia, Andrew Braunstein, Andrew Liu, Andrew Schmidt, Andrey Mereskin, Andrey Mishchenko, Andy Applebaum, Andy Rogerson, Ann Rajan, Annie Wei, Anoop Kotha, Anubha Srivastava, Anushree Agrawal, Arun Vijayvergiya, Ashley Tyra, Ashvin Nair, Avi Nayak, Ben Eggers, Bessie Ji, Beth Hoover, Bill Chen, Blair Chen, Boaz Barak, Borys Minaiev, Botao Hao, Bowen Baker, Brad Lightcap, Brandon McKinzie, Brandon Wang, Brendan Quinn, Brian Fioca, Brian Hsu, Brian Yang, Brian Yu, Brian Zhang, Brittany Brenner, Callie Riggins Zetino, Cameron Raymond, Camillo Lugaresi, Carolina Paz, Cary Hudson, Cedric Whitney, Chak Li, Charles Chen, Charlotte Cole, Chelsea Voss, Chen Ding, Chen Shen, Chengdu Huang, Chris Colby, Chris Hallacy, Chris Koch, Chris Lu, Christina Kaplan, Christina Kim, CJ Minott-Henriques, Cliff Frey, Cody Yu, Coley Czarnecki, Colin Reid, Colin Wei, Cory Decareaux, Cristina Scheau, Cyril Zhang, Cyrus Forbes, Da Tang, Dakota Goldberg, Dan Roberts, Dana Palmie, Daniel Kappler, Daniel Levine, Daniel Wright, Dave Leo, David Lin, David Robinson, Declan Grabb, Derek Chen, Derek Lim, Derek Salama, Dibya Bhattacharjee, Dimitris Tsipras, Dinghua Li, Dingli Yu, DJ Strouse, Drew Williams, Dylan Hunn, Ed Bayes, Edwin Arbus, Ekin Akyurek, Elaine Ya Le, Elana Widmann, Eli Yani, Elizabeth Proehl, Enis Sert, Enoch Cheung, Eri Schwartz, Eric Han, Eric Jiang, Eric Mitchell, Eric Sigler, Eric Wallace, Erik Ritter, Erin Kavanaugh, Evan Mays, Evgenii Nikishin, Fangyuan Li, Felipe Petroski Such, Filipe de Avila

Belbute Peres, Filippo Raso, Florent Bekerman, Foivos Tsimpourlas, Fotis Chantzis, Francis Song, Francis Zhang, Gaby Raila, Garrett McGrath, Gary Briggs, Gary Yang, Giambattista Parascandolo, Gildas Chabot, Grace Kim, Grace Zhao, Gregory Valiant, Guillaume Leclerc, Hadi Salman, Hanson Wang, Hao Sheng, Haoming Jiang, Haoyu Wang, Haozhun Jin, Harshit Sikchi, Heather Schmidt, Henry Aspegren, Honglin Chen, Huida Qiu, Hunter Lightman, Ian Covert, Ian Kivlichan, Ian Silber, Ian Sohl, Ibrahim Hammoud, Ignasi Clavera, Ikai Lan, Ilge Akkaya, Ilya Kostrikov, Irina Kofman, Isak Etinger, Ishaan Singal, Jackie Hehir, Jacob Huh, Jacqueline Pan, Jake Wilczynski, Jakub Pachocki, James Lee, James Quinn, Jamie Kiros, Janvi Kalra, Jasmyn Samaroo, Jason Wang, Jason Wolfe, Jay Chen, Jay Wang, Jean Harb, Jeffrey Han, Jeffrey Wang, Jennifer Zhao, Jeremy Chen, Jerene Yang, Jerry Tworek, Jesse Chand, Jessica Landon, Jessica Liang, Ji Lin, Jiancheng Liu, Jianfeng Wang, Jie Tang, Jihan Yin, Joanne Jang, Joel Morris, Joey Flynn, Johannes Ferstad, Johannes Heidecke, John Fishbein, John Hallman, Jonah Grant, Jonathan Chien, Jonathan Gordon, Jongsoo Park, Jordan Liss, Jos Kraaijeveld, Joseph Guay, Joseph Mo, Josh Lawson, Josh McGrath, Joshua Vendrow, Joy Jiao, Julian Lee, Julie Steele, Julie Wang, Junhua Mao, Kai Chen, Kai Hayashi, Kai Xiao, Kamyar Salahi, Kan Wu, Karan Sekhri, Karan Sharma, Karan Singhal, Karen Li, Kenny Nguyen, Keren Gu-Lemberg, Kevin King, Kevin Liu, Kevin Stone, Kevin Yu, Kristen Ying, Kristian Georgiev, Kristie Lim, Kushal Tirumala, Kyle Miller, Lama Ahmad, Larry Lv, Laura Clare, Laurance Fauconnet, Lauren Itow, Lauren Yang, Laurentia Romaniuk, Leah Anise, Lee Byron, Leher Pathak, Leon Maksin, Leyan Lo, Leyton Ho, Li Jing, Liang Wu, Liang Xiong, Lien Mamitsuka, Lin Yang, Lindsay McCallum, Lindsey Held, Liz Bourgeois, Logan Engstrom, Lorenz Kuhn, Louis Feuvrier, Lu Zhang, Lucas Switzer, Lukas Kondraciuk, Lukasz Kaiser, Manas Joglekar, Mandeep Singh, Mandip Shah, Manuka Stratta, Marcus Williams, Mark Chen, Mark Sun, Marselus Cayton, Martin Li, Marvin Zhang, Marwan Aljubeh, Matt Nichols, Matthew Haines, Max Schwarzer, Mayank Gupta, Meghan Shah, Melody Huang, Meng Dong, Mengqing Wang, Mia Glaese, Micah Carroll, Michael Lampe, Michael Malek, Michael Sharman, Michael Zhang, Michele Wang, Michelle Pokrass, Mihai Florian, Mikhail Pavlov, Miles Wang, Ming Chen, Mingxuan Wang, Minnia Feng, Mo Bavarian, Molly Lin, Moose Abdool, Mostafa Rohaninejad, Nacho Soto, Natalie Staudacher, Natan LaFontaine, Nathan Marwell, Nelson Liu, Nick Preston, Nick Turley, Nicklas Ansman, Nicole Blades, Nikil Pancha, Nikita Mikhaylin, Niko Felix, Nikunj Handa, Nishant Rai, Nitish Keskar, Noam Brown, Ofir Nachum, Oleg Boiko, Oleg Murk, Olivia Watkins, Oona Gleeson, Pamela Mishkin, Patryk Lesiewicz, Paul Baltescu, Pavel Belov, Peter Zhokhov, Philip Pronin, Phillip Guo, Phoebe Thacker, Qi Liu, Qiming Yuan, Qinghua Liu, Rachel Dias, Rachel Puckett, Rahul Arora, Ravi Teja Mullapudi, Raz Gaon, Reah Miyara, Rennie Song, Rishabh Aggarwal, RJ Marsan, Robel Yemiru, Robert Xiong, Rohan Kshirsagar, Rohan Nuttall, Roman Tsiupa, Ronen Eldan, Rose Wang, Roshan James, Roy Ziv, Rui Shu, Ruslan Nigmatullin, Saachi Jain, Saam Talaie, Sam Altman, Sam Arnesen, Sam Toizer, Sam Toyer, Samuel Miserendino, Sandhini Agarwal, Sarah Yoo, Savannah Heon, Scott Ethersmith, Sean Grove, Sean Taylor, Sebastien Bubeck, Sever Banesiu, Shaokyi Amdo, Shengjia Zhao, Sherwin Wu, Shibani Santurkar, Shiyu Zhao, Shraman Ray Chaudhuri, Shreyas Krishnaswamy, Shuaiqi, Xia, Shuyang Cheng, Shyamal Anadkat, Simón Posada Fishman, Simon Tobin, Siyuan Fu, Somay Jain, Song Mei, Sonya Egoian, Spencer Kim, Spug Golden, SQ Mah, Steph Lin, Stephen Imm, Steve Sharpe, Steve Yadlowsky, Sulman Choudhry, Sungwon Eum, Suvansh Sanjeev, Tabarak Khan, Tal Stramer, Tao Wang, Tao Xin, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Degry, Thomas Shadwell, Tianfu Fu, Tianshi Gao, Timur Garipov, Tina Sriskandarajah, Toki Sherbakov, Tomer Kaftan, Tomo Hiratsuka, Tongzhou Wang, Tony Song, Tony Zhao, Troy Peterson, Val Kharitonov, Victoria Chernova, Vineet Kosaraju, Vishal Kuo, Vitchyr Pong, Vivek Verma, Vlad Petrov, Wanning Jiang, Weixing Zhang, Wenda Zhou, Wenlei Xie, Wenting Zhan, Wes McCabe, Will DePue, Will Ellsworth, Wulfie Bain, Wyatt Thompson, Xiangning Chen, Xiangyu Qi, Xin Xiang, Xinwei Shi, Yann Dubois, Yaodong Yu, Yara Khakbaz, Yifan Wu, Yilei Qian, Yin Tat Lee, Yinbo Chen, Yizhen Zhang, Yizhong Xiong, Yonglong Tian, Young Cha, Yu Bai, Yu Yang, Yuan Yuan, Yuanzhi Li, Yufeng Zhang, Yuguang Yang, Yujia Jin, Yun Jiang, Yunyun Wang, Yushi Wang, Yutian Liu, Zach Stubenvoll, Zehao Dou, Zheng Wu, and Zhigang Wang. OpenAI GPT-5 system card, 2025. URL `https://arxiv.org/abs/2601.03267`.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Weinan Dai, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, Jinhua Zhu, Jiaze Chen, Jiangjie Chen, Chengyi Wang, Hongli Yu, Yuxuan Song, Xiangpeng Wei, Hao Zhou, Jingjing Liu, Wei-Ying Ma, Ya-Qin Zhang, Lin Yan, Mu Qiao, Yonghui Wu, and Mingxuan Wang. Dapo: An open-source llm

reinforcement learning system at scale, 2025b. URL https://arxiv.org/abs/2503.14476.

Chang Gao, Chujie Zheng, Xiong-Hui Chen, Kai Dang, Shixuan Liu, Bowen Yu, An Yang, Shuai Bai, Jingren Zhou, and Junyang Lin. Soft adaptive policy optimization, 2025. URL https://arxiv.org/abs/2511.20347.

Zichen Liu, Changyu Chen, Wenjun Li, Penghui Qi, Tianyu Pang, Chao Du, Wee Sun Lee, and Min Lin. Understanding r1-zero-like training: A critical perspective, 2025b. URL https://arxiv.org/abs/2503.20783.

Shih-Yang Liu, Xin Dong, Ximing Lu, Shizhe Diao, Mingjie Liu, Min-Hung Chen, Hongxu Yin, Yu-Chiang Frank Wang, Kwang-Ting Cheng, Yejin Choi, Jan Kautz, and Pavlo Molchanov. Dler: Doing length penalty right - incentivizing more intelligence per token via reinforcement learning, 2025c. URL https://arxiv.org/abs/2510.15110.

Kimi Team, Tongtong Bai, Yifan Bai, Yiping Bao, S. H. Cai, Yuan Cao, Y. Charles, H. S. Che, Cheng Chen, Guanduo Chen, Huarong Chen, Jia Chen, Jiahao Chen, Jianlong Chen, Jun Chen, Kefan Chen, Liang Chen, Ruijue Chen, Xinhao Chen, Yanru Chen, Yanxu Chen, Yicun Chen, Yimin Chen, Yingjiang Chen, Yuankun Chen, Yujie Chen, Yutian Chen, Zhirong Chen, Ziwei Chen, Dazhi Cheng, Minghan Chu, Jialei Cui, Jiaqi Deng, Muxi Diao, Hao Ding, Mengfan Dong, Mengnan Dong, Yuxin Dong, Yuhao Dong, Angang Du, Chenzhuang Du, Dikang Du, Lingxiao Du, Yulun Du, Yu Fan, Shengjun Fang, Qiulin Feng, Yichen Feng, Garimugai Fu, Kelin Fu, Hongcheng Gao, Tong Gao, Yuyao Ge, Shangyi Geng, Chengyang Gong, Xiaochen Gong, Zhuoma Gongque, Qizheng Gu, Xinran Gu, Yicheng Gu, Longyu Guan, Yuanying Guo, Xiaoru Hao, Weiran He, Wenyang He, Yunjia He, Chao Hong, Hao Hu, Jiaxi Hu, Yangyang Hu, Zhenxing Hu, Ke Huang, Ruiyuan Huang, Weixiao Huang, Zhiqi Huang, Tao Jiang, Zhejun Jiang, Xinyi Jin, Yu Jing, Guokun Lai, Aidi Li, C. Li, Cheng Li, Fang Li, Guanghe Li, Guanyu Li, Haitao Li, Haoyang Li, Jia Li, Jingwei Li, Junxiong Li, Lincan Li, Mo Li, Weihong Li, Wentao Li, Xinhang Li, Xinhao Li, Yang Li, Yanhao Li, Yiwei Li, Yuxiao Li, Zhaowei Li, Zheming Li, Weilong Liao, Jiawei Lin, Xiaohan Lin, Zhishan Lin, Zichao Lin, Cheng Liu, Chenyu Liu, Hongzhang Liu, Liang Liu, Shaowei Liu, Shudong Liu, Shuran Liu, Tianwei Liu, Tianyu Liu, Weizhou Liu, Xiangyan Liu, Yangyang Liu, Yanming Liu, Yibo Liu, Yuanxin Liu, Yue Liu, Zhengying Liu, Zhongnuo Liu, Enzhe Lu, Haoyu Lu, Zhiyuan Lu, Junyu Luo, Tongxu Luo, Yashuo Luo, Long Ma, Yingwei Ma, Shaoguang Mao, Yuan Mei, Xin Men, Fanqing Meng, Zhiyong Meng, Yibo Miao, Minqing Ni, Kun Ouyang, Siyuan Pan, Bo Pang, Yuchao Qian, Ruoyu Qin, Zeyu Qin, Jiezhong Qiu, Bowen Qu, Zeyu Shang, Youbo Shao, Tianxiao Shen, Zhennan Shen, Juanfeng Shi, Lidong Shi, Shengyuan Shi, Feifan Song, Pengwei Song, Tianhui Song, Xiaoxi Song, Hongjin Su, Jianlin Su, Zhaochen Su, Lin Sui, Jinsong Sun, Junyao Sun, Tongyu Sun, Flood Sung, Yunpeng Tai, Chuning Tang, Heyi Tang, Xiaojuan Tang, Zhengyang Tang, Jiawen Tao, Shiyuan Teng, Chaoran Tian, Pengfei Tian, Ao Wang, Bowen Wang, Chensi Wang, Chuang Wang, Congcong Wang, Dingkun Wang, Dinglu Wang, Dongliang Wang, Feng Wang, Hailong Wang, Haiming Wang, Hengzhi Wang, Huaqing Wang, Hui Wang, Jiahao Wang, Jinhong Wang, Jiuzheng Wang, Kaixin Wang, Linian Wang, Qibin Wang, Shengjie Wang, Shuyi Wang, Si Wang, Wei Wang, Xiaochen Wang, Xinyuan Wang, Yao Wang, Yejie Wang, Yipu Wang, Yiqin Wang, Yucheng Wang, Yuzhi Wang, Zhaoji Wang, Zhaowei Wang, Zhengtao Wang, Zhexu Wang, Zihan Wang, Zizhe Wang, Chu Wei, Ming Wei, Chuan Wen, Zichen Wen, Chengjie Wu, Haoning Wu, Junyan Wu, Rucong Wu, Wenhao Wu, Yuefeng Wu, Yuhao Wu, Yuxin Wu, Zijian Wu, Chenjun Xiao, Jin Xie, Xiaotong Xie, Yuchong Xie, Yifei Xin, Bowei Xing, Boyu Xu, Jianfan Xu, Jing Xu, Jinjing Xu, L. H. Xu, Lin Xu, Suting Xu, Weixin Xu, Xinbo Xu, Xinran Xu, Yangchuan Xu, Yichang Xu, Yuemeng Xu, Zelai Xu, Ziyao Xu, Junjie Yan, Yuzi Yan, Guangyao Yang, Hao Yang, Junwei Yang, Kai Yang, Ningyuan Yang, Ruihan Yang, Xiaofei Yang, Xinlong Yang, Ying Yang, Yi Yang, Yi Yang, Zhen Yang, Zhilin Yang, Zonghan Yang, Haotian Yao, Dan Ye, Wenjie Ye, Zhuorui Ye, Bohong Yin, Chengzhen Yu, Longhui Yu, Tao Yu, Tianxiang Yu, Enming Yuan, Mengjie Yuan, Xiaokun Yuan, Yang Yue, Weihao Zeng, Dunyuan Zha, Haobing Zhan, Dehao Zhang, Hao Zhang, Jin Zhang, Puqi Zhang, Qiao Zhang, Rui Zhang, Xiaobin Zhang, Y. Zhang, Yadong Zhang, Yangkun Zhang, Yichi Zhang, Yizhi Zhang, Yongting Zhang, Yu Zhang, Yushun Zhang, Yutao Zhang, Yutong Zhang, Zheng Zhang, Chenguang Zhao, Feifan Zhao, Jinxiang Zhao, Shuai Zhao, Xiangyu Zhao, Yikai Zhao, Zijia Zhao, Huabin Zheng, Ruihan Zheng, Shaojie Zheng, Tengyang Zheng, Junfeng Zhong, Longguang Zhong, Weiming Zhong, M. Zhou, Runjie Zhou, Xinyu Zhou, Zaida Zhou, Jinguo Zhu, Liya Zhu, Xinhao Zhu, Yuxuan Zhu, Zhen Zhu, Jingze Zhuang, Weiyu Zhuang, Ying Zou, and Xinxing Zu. Kimi k2.5: Visual agentic intelligence, 2026. URL https://arxiv.org/abs/2602.02276.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024.

Deheng Ye, Zhao Liu, Mingfei Sun, Bei Shi, Peilin Zhao, Hao Wu, Hongsheng Yu, Shaojie Yang, Xipeng Wu, Qingwei Guo, Qiaobo Chen, Yinyuting Yin, Hao Zhang, Tengfei Shi, Liang Wang, Qiang Fu, Wei Yang, and Lanxiao Huang. Mastering complex control in moba games with deep reinforcement learning, 2020. URL `https://arxiv.org/abs/1912.09729`.

A Appendix

A.1 Implementation Details

We build our codebase and implement c-GDPO on top of verl Sheng et al. [2024]. We set the number of rollouts to 8, the learning rate to 5e-6, and the batch size to 256. We use a cosine learning rate scheduler and train for 1 epoch. For main results in Table 1, we use $w_{\text{prec}} = 0.1$, $w_{\text{rec}} = 0.3$, $w_{\text{ling}} = 0.3$, $\tau_l = 0.5$, and $\tau_u = 2$ for better empirical results.

Our training data consist of the initial 100K image-text pairs of ShareGPT4V Chen et al. [2023], whose captions are annotated by GPT-4V OpenAI et al. [2024] (not the 1.2M-pair ShareGPT4V-PT set); after image de-duplication, roughly 90K image-text pairs remain. We compare the original captions with a recapped version generated by GPT-5-mini. In Table 1, GPT-4o-mini is used to decompose on-policy captions into atomic assertions and to label each assertion and also used to label assertions for whether it is correct, pointable and covered in the reference caption.

We empirically observed that removing the KL divergence loss Yu et al. [2025b] and applying dual-clip Ye et al. [2020] improves training stability (which is an implementation detail omitted from Section 3 to keep it simple). Following DAPO Yu et al. [2025b], we also adopted "`token-sum-sequence-mean`" which further improves empirical performance.

Every single experiment is conducted using one 8-gpus B-200 node. For the main results that use GPT-4o-mini as judge, it takes roughly 24 hours for LLaVA-7B experiments, 37 hours for QwenVL2.5-3B experiments and 46 hours for QwenVL2.5-7B experiments.

A.2 Training Prompt

```
You are an expert image caption evaluator. You will analyze a SYNTHETIC CAPTION and a GROUND TRUTH CAPTION
      for an image to extract detailed features for quality assessment.

SYNTHETIC CAPTION (to evaluate):
"%(caption)s"

GROUND TRUTH CAPTION (reference):
"%(g_caption)s"

Your task is to extract the following features in JSON format:

1. Atomic Assertions: Break each caption into shortest possible factual claims.
   Extract ALL assertions from the caption, including meta-commentary and subjective statements.
   Do NOT filter out any assertions during extraction - we need the complete list.

2. Precision and Recall - VERIFICATION REQUIRES BOTH CONDITIONS:

   ======================================================================
   VERIFICATION RUBRIC: TWO-PART TEST
   ======================================================================

   For EACH synthetic assertion, mark is_verified: true ONLY if BOTH conditions are met:

   CONDITION 1 - THE "POINT TO IT" TEST:
   Could a person physically POINT TO this thing in the image?

   [PASS] Physical objects, visible attributes, spatial locations, visible text
     - "a car", "a red door", "silver hubcap", "on the left", "sign reads STOP"

   [FAIL] Abstractions, effects, judgments, emotions, relationships, intentions
     - "adds depth", "contrasting nicely", "inviting the viewer", "happy", "family"

   CONDITION 2 - VISUAL VERIFICATION:
```

```
   Is this assertion actually TRUE in the image?
   Check against the ground truth caption OR examine the image directly.

   [PASS] The described object/attribute/location is actually visible in the image
   [FAIL] The assertion is a hallucination or factually incorrect

   ======================================================================

   VERIFICATION DECISION TABLE:

   | Assertion            | Pointable? | Actually in Image? | is_verified |
   |----------------------|------------|--------------------|-------------|
   | "A silver hubcap"    | Yes        | Yes (exists)       | TRUE        |
   | "A silver hubcap"    | Yes        | No (no hubcap)     | FALSE       |
   | "adds depth"         | No         | N/A                | FALSE       |
   | "contrasting nicely" | No         | N/A                | FALSE       |
   | "A purple elephant"  | Yes        | No (not there)     | FALSE       |

   KEY PRINCIPLE: An assertion that fails EITHER test must be marked is_verified: false.
   - Meta commentary fails the "point to it" test -> is_verified: false
   - Hallucinated objects fail the visual verification test -> is_verified: false
   - Only correct, pointable facts pass both tests -> is_verified: true

   For RECALL: Check if synthetic caption covers GT content (exact wording not needed)

3. Evaluate the linguistic quality of the synthetic caption, independent of its accuracy to the image.
     Assess the following:
- Clarity: How easily can a human reader understand the caption? Consider absence of ambiguous or overly
     complex phrasing, simplicity and readability, avoidance of unnecessary detail that obscures meaning.
     Penalize overly long or overloaded sentences that pack too many ideas or modifiers into a single
     sentence, making the caption difficult to parse or unnatural for human readers. Penalize unnecessary
     repetition or redundant wording that does not add new information.
- Fluency: How natural and well-formed is the writing? Consider correct grammar, punctuation, and spelling,
      natural sentence flow, human-like phrasing, not a list of facts, avoidance of unnatural compound
     words, excessive adjective stacking, or run-on sentences.
- Coherency: How well do different parts of the caption connect to form one unified statement? Consider
     logical ordering of information, smooth transitions between ideas, consistent perspective and focus,
     no abrupt topic shifts.

Based on the above criteria, rate each item (Clarity, Fluency, Coherency) from 1 to 10, where 10 is the
     highest score given for excellent, highly natural, and easy to read synthetic caption and 1 is the
     lowest score for very poor, confusing, or unnatural synthetic caption. After rating, provide a brief
     summary explaining the reasoning behind the given quality scores especially noting problems such as
     convoluted syntax, excessive modifiers, disjointed structure, unnatural or robotic phrasing, invented
      or overly technical compound words.

Return ONLY valid JSON in this EXACT format (no markdown, no code blocks):
{
    "synthetic_features": {
        "atomic_assertions": [
            {"text": str, "is_verified": bool},
            ...
        ],
        "clarity_score": int,
        "fluency_score": int,
        "coherency_score": int,
        "linguistic_scores_explanation": str
    },
    "gt_features": {
        "atomic_assertions": [
            {"text": str, "is_covered": bool},
            ...
        ]
    }
}

IMPORTANT RULES:
- EXTRACT ALL: Include ALL assertions from the caption, even meta-commentary and subjective statements.
- TWO-PART TEST: is_verified: true requires BOTH (1) pointable AND (2) actually visible in image.
- PENALIZE BOTH: Meta commentary AND hallucinations should both result in is_verified: false.
- Be strict: when in doubt about either condition, mark as NOT VERIFIED.
- For recall: check if synthetic caption covers GT content (not exact wording needed)
- Return ONLY the JSON object, no other text or formatting
```

A.3 Full Proof

Proof of Proposition 1. Consider a K-reward, G-rollout setting. We assume G >= 3 to ensure the group-wise standard deviation is well-defined and the resulting normalization is non-trivial. Let the

reward vector of rollout i be

$$\mathbf{r}_i = (r_{i,1}, \ldots, r_{i,K}) \in \mathbb{R}^K,$$

and let $w_k \in \mathbb{R}$ denote the aggregation weight for reward dimension k.

Vanilla GRPO. Vanilla multi-reward GRPO first aggregates rewards into a single scalar

$$s_i = \sum_{k=1}^{K} w_k r_{i,k}.$$

It then computes the group-normalized advantage from the scalar scores $\{s_i\}_{i=1}^G$:

$$A_i^{\mathrm{GRPO}} = \frac{s_i - \mu_s}{\sigma_s}, \qquad \mu_s = \frac{1}{G}\sum_{j=1}^{G} s_j, \qquad \sigma_s = \sqrt{\frac{1}{G}\sum_{j=1}^{G}(s_j - \mu_s)^2},$$

where we use the population standard deviation for concreteness. The same argument holds under sample normalization up to a constant factor.

Fix rollout $i = 1$, and fix the competing rollouts $\mathbf{r}_2, \ldots, \mathbf{r}_G$. Then $s_2, \ldots, s_G$ are fixed constants, and the only variable is $\mathbf{r}_1$. Since

$$s_1 = \sum_{k=1}^{K} w_k r_{1,k},$$

both μ_s and σ_s are functions of s_1 only:

$$\mu_s = \frac{s_1 + \sum_{j=2}^{G} s_j}{G}, \qquad \sigma_s = \sigma_s(s_1; s_2, \ldots, s_G).$$

Therefore there exists a scalar function $f : \mathbb{R} \to \mathbb{R}$, depending on the fixed competing rollouts, such that

$$A_1^{\mathrm{GRPO}} = f(s_1) = f\left(\sum_{k=1}^{K} w_k r_{1,k}\right).$$

Hence the normalized advantage of rollout 1 depends on its reward vector $\mathbf{r}_1$ only through the scalar weighted sum $\sum_{k=1}^{K} w_k r_{1,k}$.

Now consider two reward vectors $\mathbf{r}_1, \mathbf{r}_1' \in \mathbb{R}^K$ such that

$$\sum_{k=1}^{K} w_k r_{1,k} = \sum_{k=1}^{K} w_k r'_{1,k}.$$

Then $s_1 = s_1'$, which implies

$$A_1^{\mathrm{GRPO}}(\mathbf{r}_1) = f(s_1) = f(s_1') = A_1^{\mathrm{GRPO}}(\mathbf{r}_1').$$

Therefore, all reward vectors lying on the same weighted-sum hyperplane

$$\left\{ \mathbf{r} \in \mathbb{R}^K : \sum_{k=1}^{K} w_k r_k = c \right\}$$

are indistinguishable to vanilla GRPO. This proves that reward aggregation before group normalization induces a many-to-one mapping from $\mathbb{R}^K$ to scalar normalized advantages, and that optimization depends only on the aggregated reward direction while discarding orthogonal reward-trade-off information.

c-GDPO. In reward-decoupled normalization, each reward dimension is normalized separately before aggregation. Define

$$\mu_k = \frac{1}{G}\sum_{j=1}^{G} r_{j,k}, \qquad \sigma_k = \sqrt{\frac{1}{G}\sum_{j=1}^{G}(r_{j,k} - \mu_k)^2},$$

and let the normalized per-reward deviation of rollout i be

$$\widetilde{A}_{i,k} = \frac{r_{i,k} - \mu_k}{\sigma_k}.$$

The final decoupled advantage is

$$A_i^{\text{c-GDPO}} = \sum_{k=1}^{K} w_k \widetilde{A}_{i,k} = \sum_{k=1}^{K} w_k \frac{r_{i,k} - \mu_k}{\sigma_k}.$$

Again fix rollout $i = 1$ and competing rollouts $\mathbf{r}_2, \ldots, \mathbf{r}_G$. Then for each reward dimension k, the quantities μ_k and σ_k depend on $r_{1,k}$ and the fixed values $r_{2,k}, \ldots, r_{G,k}$, but do not depend on any other coordinate $r_{1,\ell}$ with $\ell \neq k$. Hence there exist scalar functions $f_k : \mathbb{R} \to \mathbb{R}$ such that

$$A_1^{\text{c-GDPO}} = \sum_{k=1}^{K} w_k f_k(r_{1,k}).$$

Therefore the decoupled advantage depends on the reward coordinates separately rather than only through the aggregated sum $\sum_{k=1}^{K} w_k r_{1,k}$.

Consequently, reward-decoupled normalization is generally not invariant to the weighted-sum hyperplanes above. In particular, suppose there exist two reward dimensions $p \neq q$ with nonzero weights w_p, w_q, and nondegenerate competing-rollout variances in those dimensions so that f_p and f_q are not constant. Consider two reward vectors $\mathbf{r}_1, \mathbf{r}_1'$ that differ only in coordinates p and q and satisfy

$$w_p r_{1,p} + w_q r_{1,q} = w_p r'_{1,p} + w_q r'_{1,q},$$

with $r_{1,p} \neq r'_{1,p}$ and $r_{1,q} \neq r'_{1,q}$. Then $\mathbf{r}_1$ and $\mathbf{r}_1'$ lie on the same weighted-sum hyperplane, but

$$A_1^{\text{c-GDPO}}(\mathbf{r}_1) - A_1^{\text{c-GDPO}}(\mathbf{r}_1') = w_p\big[f_p(r_{1,p}) - f_p(r'_{1,p})\big] + w_q\big[f_q(r_{1,q}) - f_q(r'_{1,q})\big],$$

which is generically nonzero because the two coordinates are normalized independently. Thus equal aggregate reward does not in general imply equal decoupled advantage.

Therefore, unlike vanilla GRPO, reward-decoupled normalization does not collapse all reward trade-offs sharing the same aggregated reward into the same update signal. Instead, it preserves per-reward relative deviations before aggregation, retaining finer-grained optimization information for continuous multi-reward trade-offs.

A.4 Evaluation Prompt (b-CapScore)

Same to the training prompt, but slight differences in the precision verification part to allow generalization. More specifically:

```
======================================================================
VERIFICATION RUBRIC: TWO-PART TEST
======================================================================

For EACH synthetic assertion, mark is_verified: true ONLY if BOTH conditions are met:

CONDITION 1 - THE "POINT TO IT OR POINT TO EVIDENCE THAT JUSTIFIES IT" TEST:
Could a person either:
(1) physically POINT TO the asserted thing in the image, OR
(2) point to visible evidence that reasonably JUSTIFIES the assertion using ordinary world knowledge?

 PASSES - Directly Visible Facts:
  - "a car"
  - "a red door"
  - "silver hubcap"
  - "on the left"
  - "sign reads STOP"

 PASSES - Visually Justified Inferences Using World Knowledge:
  - "the road is wet": puddles, reflections, surface sheen
  - "the building is under construction": scaffolding, unfinished surfaces, exposed structure
  - "the person is a bride": wedding dress, veil, bouquet, ceremony context
  - "the man is likely a chef": chef hat, apron, kitchen setting
  - "the child is blowing out birthday candles": cake, candles, leaning posture
```

```
  - "the people are waiting in line": queue formation toward a counter or entrance
  - "the room is set up for a meeting": conference table, arranged chairs, screen, notebooks
  - "the person is reading a menu": menu-like object, restaurant setting, gaze/posture

These pass because the claim is not directly visible as a single object, but the image provides clear
     evidence that justifies the inference using ordinary shared world knowledge.

 FAILS - Unsupported Speculation:
  - "adds depth to the composition"
  - "contrasting nicely"
  - "inviting the viewer"
  - "creates a welcoming atmosphere"
  - "the photographer intended to..."
  - "the person feels proud"
  - "this suggests luxury" (without very strong visible evidence)
  - "they are a family" (unless the image gives unusually strong evidence)
  - "this is a reunion"
  - "the person is thinking about leaving"

IMPORTANT:
World knowledge IS allowed when it is used to make a normal, evidence-based visual inference.
World knowledge is NOT allowed when it becomes mind-reading, storytelling, symbolism, or weak social
     speculation.

CONDITION 2 - VISUAL VERIFICATION:
Is this assertion actually TRUE in the image?

Use the image as the PRIMARY source of truth.
The ground-truth caption may help, but do not reject a claim only because it is not mentioned in the
     ground-truth caption.

 PASSES:
  - the assertion is visually supported and actually true

 FAILS:
  - the assertion is hallucinated
  - the assertion is factually incorrect
  - the evidence is too weak or ambiguous to support the inference confidently

======================================================================

KEY PRINCIPLE:
An assertion should be marked is_verified: true only if it is either directly visible OR justified by
     visible evidence through ordinary world knowledge, AND it is actually true for this image.
```

A.5 A Balanced Captioning Metric (b-CapScore)

While we advocate to use multiple captioning benchmarks to evaluate the quality of the captioning, we show that it is possible to turn our reward to one metric to evaluate the captioning quality in a more balanced way.

We reuse the images and human reference captions of DCScore Ye et al. [2025] as the data sources and define the balanced metric b-CapScore as the harmonic mean of pointability-aware precision, reference coverage, and linguistic quality:

$$\text{b-CapScore} = \frac{3}{\frac{1}{R_{\text{prec}}} + \frac{1}{R_{\text{rec}}} + \frac{1}{R_{\text{ling}}}}. \tag{8}$$

Here, R_{prec} is computed using the same pointability-aware verification rubric as in training: an atomic assertion is counted as correct only if it is visually verified and pointable, or supported by pointable visual evidence (full prompt in Appendix A.4). R_{rec} measures reference coverage, and R_{ling} denotes the normalized linguistic-quality score. By using a harmonic mean, b-CapScore penalizes imbalance across correctness, coverage, and linguistic quality. To ensure we reduce the error from atomic assertion decomposition, we use GPT-5.4 as the judge for b-CapScore evaluation for Table 1.

Table 6: Comparison on Model-level Spearman

Metric	Ranking Procedure	Model-level Spearman
GPT-4o-as-a-Judge w/ ref	Arena/ELO from pairwise judgments	0.943
DCScore	Average score over captions	0.943
b-CapScore	Average score over captions	**0.956**

To better understand the proposed balanced captioning metric, we perform human alignment analysis on CapArena and compare the human alignment between CapArena and b-CapScore in Table 6. CapArena reports model-level Spearman by converting metric judgments into pairwise arena outcomes and deriving model rankings through the same arena/ELO-style procedure used for human preferences. In contrast, our metric does not require arena-style pairwise comparison: for each model, we compute the average reference-based caption score over the evaluation set, rank models by this average score, and report Spearman's rank correlation with the human-derived CapArena ranking. This shows that our reward can be turned to a metric that has comparable or better human alignment than arena-style comparison.

A.6 Reward Weight Ablation Studies

Table 7: Results for different values of w_{prec}.

w_{prec}	DCScore	CaptionQA	CapArena
0.0	41.2	73.6	-13.8
0.1	50.8	75.0	-3.8
0.2	52.8	74.9	-5.8
0.3	52.0	75.0	-12.0
0.4	52.1	75.3	-10.0

In Table 7, we analyze sensitivity to the reward weights by varying w_{prec} while keeping the other two reward weights fixed. Setting $w_{\text{prec}} = 0$ leads to the worst overall performance. For non-zero values of w_{prec}, increasing its weight introduces trade-offs across benchmarks.

A.7 Impact of training-time MLLM judge

Table 8: Comparison of using different training-time judge.

Model	**DCScore**	**CaptionQA**	**CapArena**	**Arena Length**
3B models				
QwenVL2.5-3B	43.3	70.0	-34.0	131
BalCapRL-3B w/ GPT-4o-mini	50.8	75.0	**-3.8**	175
BalCapRL-3B w/ GPT-5-mini	**50.9**	75.4	-6.3	186
BalCapRL-3B w/ GPT-5.4	50.8	**75.9**	-5.2	163
7B models				
QwenVL2.5-7B	46.0	74.9	13.7	136
BalCapRL-7B w/ GPT-4o-mini	53.4	79.1	28.5	192
BalCapRL-7B w/ GPT-5-mini	52.2	80.8	35.3	191
BalCapRL-7B w/ GPT-5.4	**54.2**	**81.1**	**41.7**	172

A.8 Additional Qualitative Examples



Figure 4: Additional qualitative example.

Human Reference Caption

A medium shot of a large, oval pumpkin with a carved face at night. The pumpkin is orange, with indented ridges running from the top to the bottom and a short, thick stem cut off at the top. The pumpkin has one triangle eye carved at the top center with a big open month [sic] smile carved below it. One small tooth on the top left side is part of the carved smile. The glow inside the pumpkin is golden. A small amount of dried leaves and dirt is visible inside at the bottom. An orange light coming from the left shines on the left side of the pumpkin and illuminates the green and dried plant life around it. A thick cord is visible in the background, horizontally behind the pumpkin.

QwenVL2.5-3B

The image shows a pumpkin that has been carved into a jack-o'-lantern. The pumpkin is illuminated from within, casting a warm glow. The carving features a simple face with a triangular nose and a small, round eye. The mouth is not visible in this view. The pumpkin is placed outdoors, surrounded by greenery and fallen leaves, indicating it might be in a garden or a park. The lighting suggests it is either dusk or nighttime, as there is a red hue to the overall image, possibly due to artificial lighting or the time of day. The pumpkin appears to be part of a Halloween decoration or event.

CapRL-3B

The image depicts a creatively carved pumpkin, likely intended for Halloween decoration. Here is a detailed description: 1. **Object and Setting:** - The image shows a pumpkin that has been carved into a jack-o'-lantern, specifically designed to resemble a character with a yellow face. 2. **Pumpkin Description:** - The pumpkin is orange, typical of a standard Halloween pumpkin, and it appears to be a large, mature variety, possibly a pumpkin used for carving. 3. **Carving Details:** - The pumpkin has a triangular yellow triangle cut out for the top of the head, resembling a hat or hair, which is red and textured, likely made from paper or cardboard. This suggests the pumpkin might be styled as a character with hair, possibly a duck or a similar figure. 4. **Face Features:** - The pumpkin's face is yellow inside, illuminated by light, likely from an internal candle or LED light, given the red hue in the image. It features: - A single yellow triangle for the eyes, positioned at the top center. - An orange leaf on the right side, likely representing a nose or another facial feature, also cut from paper or cardboard. 5. **Body and Structure: - The pumpkin has a ribbed texture typical of pumpkins, with vertical ridges running down its body. 6. **Lighting:** - The image has a red tint, possibly due to a red light source, likely from a fire or a Halloween-themed lighting setup, common during the fall season. 7. **Location and Surroundings:** - The pumpkin is situated outdoors, surrounded by foliage with autumn leaves, indicating it is likely in a garden or park during fall. The ground is covered with fallen leaves, predominantly red, suggesting it is autumn. 8. **Leaves and Plants: - The pumpkin is surrounded by green plants and bushes with red leaves, typical of fall. There are some plants with thin stems and leaves, possibly small shrubs or vines. 9. **Additional Elements: - On the right side of the pumpkin, there is a small piece of paper or cardboard with the number "4" visible, possibly part of a shopping basket or a label, indicating this might be in a residential area or a community event. 10. **Position and Orientation: - The pumpkin is upright and facing forward, with its mouth open, showing the interior filled with some leaves, likely from the carving process. This detailed description should(truncated)

RubiCap-3B

The image depicts a creatively carved pumpkin, likely intended for Halloween decoration. The pumpkin has a traditional jack-o'-lantern face, with a triangular nose and a wide, open mouth that is filled with small, yellowish-orange leaves. The top of the pumpkin features a red, ribbed texture, resembling a hat or a hood, which adds an artistic touch to the carving. The pumpkin is placed outdoors, surrounded by fallen leaves and other foliage, suggesting it is in a garden or a yard during autumn. The lighting appears to be artificial, possibly from a nearby light source, casting a warm glow on the pumpkin and highlighting its features. The background is slightly blurred, focusing attention on the pumpkin itself. The overall scene conveys a festive and seasonal atmosphere, typical of Halloween celebrations.

BalCapRL-3B

In the image, a pumpkin is prominently featured, illuminated by a warm, reddish light that casts a glow on its surface. The pumpkin has been carved to resemble a face, with a triangular nose made of a yellow material and a small, round eye cutout. The rest of the face is filled in with a dark orange or brown color, giving it a somewhat eerie appearance. The pumpkin is placed on a bed of fallen leaves, which are scattered around it, adding to the autumnal theme. The background is slightly blurred but shows more leaves and some greenery, suggesting that the pumpkin is outdoors, possibly in a garden or a park. The lighting creates a dramatic effect, highlighting the texture of the pumpkin's skin and the contrast between the carved features and the surrounding foliage.



Figure 5: Additional qualitative example 2.

Human reference caption

A front view of a stone statue that depicts a man from the chest up, turning his head to the right and his lip on his right shoulder. The man's nose is slightly above the shoulder, leaving a gap between the nose and the shoulder. The statue has patches of black dirt all over it. A black and gray bird is sat on top of the statue's head, facing toward the right of the image. A stone hand that is partially cut from the left of the image is reaching toward the bird. Multiple leaves from a tree are covering the background behind the statue. Sunlight is shining on the leaves.

QwenVL2.5-3B

The image shows a bird perched on top of a statue's head. The statue appears to be of a human figure, possibly a woman, with detailed facial features and a serene expression. The bird is positioned near the top of the statue's head, looking upwards. The background consists of green foliage, suggesting that the statue is located in an outdoor setting, possibly a park or garden. The bird has a mix of brown and white feathers, and its wings are slightly spread out. The statue itself is weathered, indicating it has been there for some time.

CapRL-3B

The image depicts a scene in what appears to be a park or garden, likely a historical or possibly a cemetery, given the presence of an old statue and a pigeon perched on it. Here is a detailed description: 1. **Setting and Subject:** - The image features an old, weathered statue of a human bust, likely from classical or ancient times, possibly Roman or Greek, given its style. The statue is carved from stone and shows signs of age and wear, with a rough, textured surface covered in dirt and moss, suggesting it is outdoors and possibly in a graveyard or a historical garden. 2. **Statue:** - The statue is of a person with a head that is prominently displayed in the foreground. The statue is of a man with a profile view, facing to the left. He has a rugged, aged face with a beard and mustache, characterized by a bald head with a brown spot, likely a birthmark or mark of age. The statue's hands are clasped together, holding something, though only part of one arm is visible, showing a stone arm with small black spots, possibly ants or insects. 3. **Pigeon: Perched on the Statue:** - A pigeon is perched on the top of the statue's head, specifically on the male figure's forehead. The pigeon has a brownish plumage with white markings on its chest and wings, typical of a rock dove or a common pigeon. 4. **Hand in the Background:** - On the left side of the image, there is a hand wearing gloves, likely belonging to a gardener or someone working in the area. The hand is raised, pointing towards the pigeon, possibly indicating that the person is either feeding the pigeon or gesturing at it. 5. **Environment:** - The background is lush with green foliage, likely trees or bushes, suggesting a park or forested area. The trees are dense and leafy, indicating it might be summer or spring due to the bright sunlight filtering through the leaves. 6. **Details of the Statue: - The statue's head is slightly tilted, and the face appears to be aged, with visible cracks and spots, typical of outdoor stone statues exposed to elements over time. 7. **Objects in the Image: - The statue is holding a stone object in its right arm, which could be a book or another item, but it is not clearly visible. 8. **Text and Numbers:** - There are no ..

RubiCap-3B

The image depicts an outdoor scene featuring a weathered statue of a person. The statue appears to be made of stone or a similar material, showing signs of age and wear, including patches of discoloration and possible damage. The statue is positioned in a natural setting, surrounded by lush green foliage, indicating that it might be located in a park or garden. On top of the statue's head, there is a bird perched, possibly a pigeon or dove, given its small size and the typical appearance of such birds. The bird is facing away from the camera, looking towards the background. In the foreground, on the left side of the image, there is a human hand reaching out towards the statue. The hand is partially visible, with fingers extended as if trying to touch or interact with the statue. The skin tone of the hand suggests it belongs to a person who is not part of the statue itself. The overall atmosphere of the image is serene and somewhat contemplative, blending elements of nature and human artistry. The juxtaposition of the aged statue and the living bird creates a contrast between the static and dynamic aspects of the scene.

BalCapRL-3B

In the image, a pigeon is perched on the top of a statue's head. The statue appears to be made of stone or concrete and features a human face with a somewhat weathered and aged appearance. The pigeon has a mix of brown and white feathers, with its wings slightly spread out. The statue's head is slightly tilted to the side, and the pigeon is positioned near the top of the head, close to the ear area. The background is lush with green foliage, indicating that the statue is likely located in an outdoor setting, possibly a park or garden. There is a hand visible in the foreground, wearing gloves, which seems to be reaching towards the pigeon. The hand is positioned to the left of the statue, and the fingers are extended as if trying to interact with the pigeon. The statue's shoulders and part of its upper body are also visible, showing more of the weathered texture of the stone.

A.9 Linear length penalty.

$$\rho = \frac{\ell_{\text{pred}}}{\ell_{\text{ref}}}, \tag{9}$$

we define the deviation from the acceptable range $[\tau_l, \tau_u]$ as

$$d(\rho) = \max(\rho - \tau_u, 0) + \max(\tau_l - \rho, 0). \tag{10}$$

The linear length penalty is then

$$p_{\text{len}} = \lambda_{\text{len}}\, d(\rho), \tag{11}$$

where λ_{len} controls the penalty strength. This penalty is subtracted from the normalized advantage:

$$\tilde{A} = A - p_{\text{len}}, \tag{12}$$

where A is the normalized advantage and $\tilde{A}$ is the penalized advantage used for optimization.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In the abstract, we discussed the main contributions and the main results.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have a limitation section to discuss the main limitations.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: To show that the issues pointed out and mitigated by GDPO are also critical for continuous-valued rewards, we show a proof of this proposition in the Appendix.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We disclosed the training data, the base model, the implementation details (in Appendix) that are sufficient to reproduce our results. We also plan to release code, data and models.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: We do not release the code and data in this submission. However, we do plan to release the data and code upon acceptance. Since the algorithm is relatively straightforward to implement using verl and the training data can be easily reproduced by recaptioning ShareGPT4V with its original prompts, there is also no major obstacle to reproduce our results.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: We made it clear which dataset we used and which evaluation benchmarks we used across all the tables.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We do not repeat experiments over multiple random seeds due to the cost of running the RL experiments.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provided the details about the type of GPU resources we used and the wall-time for completing the training for each type of model we reported in the Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have confirmed that we follow the code of ethics of NeurIPS.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: The paper is a foundational research and not tied to a particular applications.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.

- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [No]

    Justification: The paper is closer to a foundational research.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have internally properly credited and respected all the code, data and model weights. These will be well reflected too when we release our code, data and weights.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: They are well documented and will be reflected when we release the new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No crowdsourcing involved.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No human subjects involved.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: We have followed the LLM policy in the NeurIPS handbook.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.