MOMENTUM-COUPLED RUBRIC ADAPTATION FOR DETAILED IMAGE CAPTIONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Detailed image captioning requires accurate and comprehensive descriptions of fine-grained visual content, yet caption quality spans factual accuracy, information coverage, and clarity. Compared with conventional methods that rely mainly on high-quality supervision or holistic rewards, rubric-based reinforcement learning decomposes these requirements into explicit criteria and provides targeted, structured feedback. However, existing methods often use separate models for caption generation, rubric construction, and judging, which may lead to inconsistent interpretations across roles. Some dynamic rubric methods alternate updates between the caption policy and rubric generator while keeping the judge fixed, but staged optimization may still leave rubric construction and judging out of step with policy optimization. We propose MoCo Rubric, a two-stage framework that coordinates these roles. First, role-conditioned, shared-parameter multi-task supervised fine-tuning equips a single vision–language model to serve as the Caption Policy, Rubric Generator, and Rubric Judge. Then, the Generator constructs rubrics online from captions sampled by the current Policy, reference captions, and image evidence. The Judge provides rubric-based rewards, and only the Policy receives GRPO updates. As Policy updates change the candidates being evaluated, we use an exponential moving average of the Policy parameters to update one momentum model shared by the Generator and Judge. This gradual transfer lets both rubric roles track Policy updates without separate RL optimization while smoothing parameter changes that could disrupt their rubric capabilities under direct synchronization. Under bounded Policy updates, we derive an upper bound on the parameter gap between the Policy and momentum model and analyze the smoothing effect of momentum updates. Across five captioning benchmarks, MoCo Rubric achieves an average pairwise win rate of 72.83%, the best mean rank in blind ranking, and the highest average score in caption-based question answering.

1 INTRODUCTION

Detailed image captioning requires accurate and comprehensive descriptions of fine-grained visual content, including objects, attributes, actions, and relationships. It has various applications, such as providing rich supervision for vision–language aligning and training (Chen et al., 2024), helping blind people understand their captured surroundings (Gurari et al., 2020), and enabling language models to answer visual questions only using descriptions (Xing et al., 2025). However, considering the fact that one image can admit many valid descriptions and limited reference captions cannot cover every relevant details, one of central challenges is how to involve more reliable feedback on factual accuracy and coverage of important visual information beyond limited references.

Conventional image captioning methods often rely on high-quality supervised data (Chen et al., 2024; Lei et al., 2026) or holistic rewards based on caption scores and downstream question-answering utility (Rennie et al., 2017; Xing et al., 2026) to provide criterion-level yet limited feedback on factual errors and missing visual information. Very recent Rubric-based reinforcement learning methods instead involve additional feedback by decomposing caption quality into explicit, image-specific criteria and criterion-level judgments into structured rewards (Huang et al., 2026). The rewards are formed through a three-role feedback loop, where a Caption Policy produces candidates, a Rubric Generator defines the criteria, and a Rubric Judge applies them. Yet the differences from different backbones and optimization trajectories may lead to semantic inconsistency

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

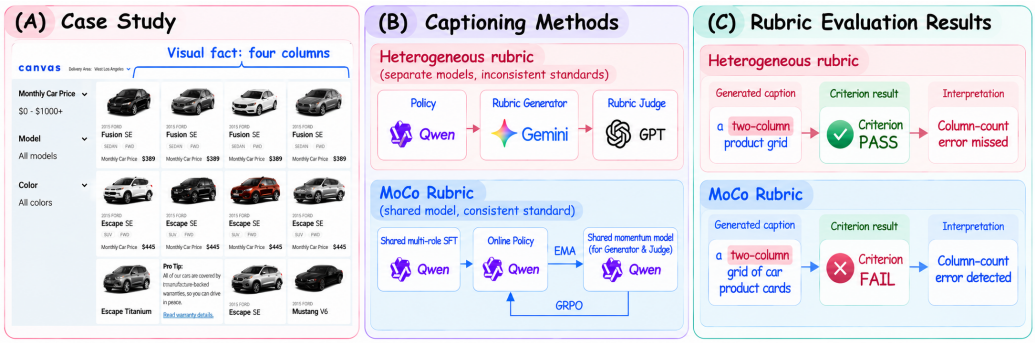


Figure 1: Comparison of heterogeneous and shared model distributions in rubric-based caption optimization.

and missing temporal coordination within the feedback loop. Firstly, the semantic inconsistency usually arises from the implementation of three roles instantiated as different backbones (Huang et al., 2026; Wang et al., 2026), such as Rubric Generator using Gemini and Rubric Judge using GPT in Fig. 1 (a). For example, Fig. 1(a) shows four columns of car listings, while a candidate caption describes a two-column grid. The Generator defines a broad criterion that checks only whether the caption mentions a grid of cars, and the Judge marks it as satisfied, leaving the column-count error undetected. Recently, EvoLM (Wang et al., 2026; Li et al., 2026) alternately optimized the Caption Policy and Rubric Generator using same or different backbones to ease such semantic inconsistency, yet the fixed Rubric Judge leads to another challenge of missing temporal coordination. The Rubric Generator might overlook important quality differences among the Policy's current candidates, and the Rubric Judge may interpret or apply the resulting criteria differently than intended. Thus, there occurs a nature question that how rubric construction and judging remain semantically consistent and temporally coordinated with an evolving Caption Policy?

Targeted it, we propose MoCo Rubric, a momentum-coupled rubric adaptation framework for detailed image captioning. It includes a Shared Multi-Role SFT stage to semantically align the three roles and a Policy RL with Momentum Transfer stage only optimizing the Caption Policy and gradually transferring its updates to two rubric roles. Specifically, the Shared Multi-Role SFT targets semantic consistency by combining a common backbone with explicit role-specific supervision. Assigning all three roles to a common backbone provides a natural starting point for reducing cross-model semantic variation by placing them in the same visual–language representation space. Sharing a backbone alone, however, does not establish the distinct capabilities required to generate captions, formulate evaluation criteria, and apply those criteria reliably. We therefore employ role-conditioned multi-task supervised fine-tuning on a single vision–language model, explicitly supervising caption generation, rubric construction, and rubric judging. The resulting checkpoint initializes both the online Caption Policy and the momentum model shared by the Rubric Generator and Rubric Judge, ensuring that all three roles enter reinforcement learning from the same learned foundation.

Further, we optimize only the Caption Policy with online rubric rewards while updating the model shared by the Rubric Generator and Rubric Judge through momentum transfer, allowing rubric feedback to adapt to the evolving Policy. Temporal coordination requires the generated rubric content responsive to predicted candidates and the Rubric Generator and Rubric Judge coordinated in parameter space within the same iteration. Existing dynamic rubric methods such as DynamicRubric and EvoLM address content adaptation by optimizing the Policy and Generator in alternating stages while keeping the Judge fixed (Wang et al., 2026; Li et al., 2026). Such stage-wise alternating optimization refreshes the Generator only at update boundaries, allowing it to lag behind the evolving Policy within each stage, while the Judge remains outside the adaptation process. Thus, our MoCo rubrics instead combines candidate-conditioned rubric construction with joint momentum transfer to the Generator and Judge, addressing both content responsiveness and evaluator-state coordination without separate reinforcement learning for either rubric role. During training, current Policy candidates, reference captions, and image evidence guide rubric construction, the Judge scores the candidates against the shared rubric, and only the Policy is optimized through GRPO (Shao et al., 2024). Because Stage I places all three roles in a shared parameter space, Policy updates may con-

tain changes to visual and linguistic representations relevant across the three tasks. These updates nevertheless optimize caption generation rather than evaluator quality, so directly copying the latest Policy parameters could disrupt the learned rubric capabilities. We therefore use an exponential moving average, inspired by MoCo (He et al., 2020), to transfer Policy changes gradually to the momentum model shared by the Generator and Judge. Both roles use the same lagged parameter version, which remains fixed throughout rubric construction and scoring for each candidate group, thereby coordinating criterion formulation and execution without separate optimization. We further analyze how the momentum coefficient and synchronization interval jointly control parameter tracking and update smoothness.

Using Qwen3-VL-8B-Instruct (Bai et al., 2025) for all three roles, MoCo Rubric achieves a 72.83% average pairwise win rate across five detailed image captioning benchmarks. It outperforms RubiCap (Huang et al., 2026) by 2.90 percentage points, even though RubiCap employs proprietary models as its Rubric Generator and Rubric Judge. MoCo Rubric also obtains the best mean rank in blind ranking and the highest average score in caption-based question answering, demonstrating improvements in both overall caption quality and downstream utility.

In summary, we introduce MoCo Rubric, a momentum-coupled rubric adaptation framework for detailed image captioning. Shared Multi-Role SFT stage uses a common backbone to establish a shared semantic foundation for the Caption Policy, Rubric Generator, and Rubric Judge. Then, we develop Policy RL with Momentum Transfer, which combines candidate-conditioned rubric construction with joint exponential-moving-average transfer of Policy updates to both rubric roles while optimizing only the Caption Policy. Further we analyze how the momentum coefficient and synchronization interval affect parameter tracking and update smoothness. Finally, we validate MoCo Rubric on five detailed image captioning benchmarks through pairwise caption comparison, blind ranking, and caption-based question answering, with ablations supporting the complementary effects of Shared Multi-Role SFT, candidate-conditioned rubrics, and joint momentum transfer.

2 Related Work

Image Captioning and Evaluation. Captioner training combines improved supervision with direct optimization of generation quality. Synthetic captioning, data filtering, and model collaboration provide rich descriptions for supervised learning (Li et al., 2022; Chen et al., 2024; Singla et al., 2024; Lei et al., 2026). RL has progressed from optimizing sequence-level metrics (Rennie et al., 2017) to rewards tailored to detailed descriptions. Recent methods assess question-answering utility (Xing et al., 2026; Yang et al., 2026a), balance factual correctness, coverage, and linguistic quality (Tang et al., 2026; Ye et al., 2026), or express image-specific requirements through rubrics (Huang et al., 2026). The approaches extend training beyond imitation of reference captions, making the reward's ability to distinguish evolving policy outputs important. Caption evaluation has expanded from reference matching to semantic and preference-based assessment. BLEU and CIDEr measure agreement with reference text (Papineni et al., 2002; Vedantam et al., 2015), while semantic and cross-modal metrics capture content correspondence (Anderson et al., 2016; Hessel et al., 2021). Fine-grained measures examine hallucinations and visual details that aggregate similarity can overlook (Rohrbach et al., 2018; Dong et al., 2024). For increasingly open-ended descriptions, LLM and VLM judges offer broader quality assessments (Chan et al., 2023; Lee et al., 2024). CapArena aggregates anonymous human pairwise preferences into model rankings, and its automated counterpart uses VLM comparisons (Cheng et al., 2025). These complementary tools assess caption quality externally; we focus on adapting structured rewards during training.

Rubric-Based Reinforcement Learning. Rubric-based RL decomposes open-ended quality requirements into explicit criteria and aggregates their judgments into rewards (Gunjal et al., 2025; Huang et al., 2025; Viswanathan et al., 2025). Beyond deriving criteria from instructions and references, rubric construction uses response contrasts, high-quality examples, and recursive refinement to improve specificity and discrimination (Liu et al., 2026; Zhang et al., 2026; Shen et al., 2026). Online approaches condition evaluation criteria on current responses, external evidence, or previously collected rubrics, allowing evaluation content to adapt as the policy changes (Jia et al., 2026; Shao et al., 2026; Guan et al., 2026; Wang et al., 2026). Building on explicit, adaptive criteria, we ask whether the parameters that generate and apply them can be driven by Policy RL alone, and examine the reliability and cost of doing so. Training rubric roles extends adaptation from evaluation content

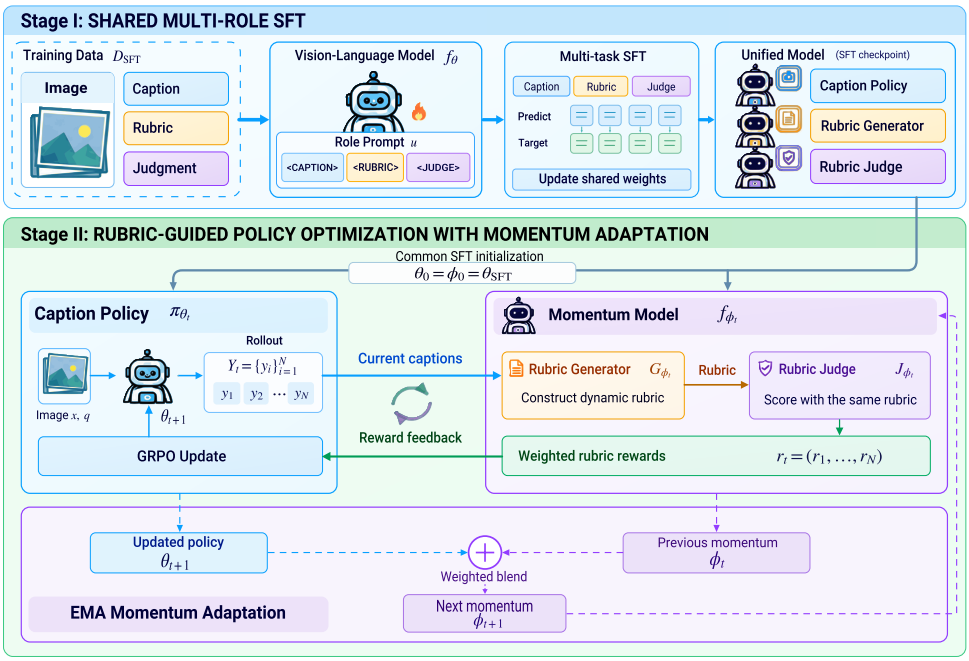


Figure 2: Overview of MoCo Rubric. Stage I trains the three roles through shared multi-role SFT. In Stage II, current Policy candidates inform the rubric, the momentum Generator and Judge assign rewards, and only the Policy receives RL gradients. The shared rubric model is updated through EMA at synchronization boundaries.

to evaluation capability. Rubric-ARM alternates generator and judge optimization (Xu et al., 2026), whereas EvoLM and DynamicRubric alternate policy and generator updates with a fixed judge or verifier (Li et al., 2026; Wang et al., 2026). Other approaches jointly optimize response and rubric generation through shared models or separate role adapters (Sheng et al., 2026; Guan et al., 2026; Ding et al., 2026). These methods introduce learning objectives for the rubric roles to improve evaluation alongside generation. Our approach starts from shared multi-role initialization, applies RL gradients only to the Policy, and updates one momentum model shared by both rubric roles through an exponential moving average of Policy parameters. This parameter-transfer mechanism removes separate RL updates and associated supervision construction for the rubric roles.

3 Methodology

3.1 Framework Overview

Given an image x and a captioning instruction q, detailed image captioning aims to generate a description y that is factually accurate and covers important visual information. Because an image can admit multiple valid descriptions, a limited set of reference captions may not capture every relevant detail. We therefore train with image-specific rubric feedback that evaluates generated captions at the criterion level. We denote the Caption Policy by π_θ, the Rubric Generator by G_ϕ, and the Rubric Judge by J_ϕ; the latter two are the rubric roles. Reference captions are used for rubric construction during training, whereas inference uses only x and q and runs the Caption Policy alone.

This feedback depends on the Policy describing the image, the Generator using it to construct rubrics, and the Judge applying those rubrics to captions. Separately instantiated roles can disagree about caption quality, and Policy updates can change the candidates that the rubric roles need to evaluate. As illustrated in Fig. 2, MoCo Rubric addresses these challenges in two stages. *Shared Multi-Role SFT* explicitly teaches one vision–language model to generate captions, construct rubrics, and judge captions under different role prompts. *Policy RL with Momentum Transfer* then uses current Policy candidates to construct rubrics and optimize only the Policy; an exponential moving average

(EMA) transfers its parameter updates to a single momentum model shared by the Generator and Judge. These two stages establish a common role foundation and keep rubric feedback responsive as the Policy evolves.

3.2 SHARED MULTI-ROLE SFT

Using the same backbone for all three roles provides a common starting point, but does not by itself teach the model how to formulate or apply evaluation criteria. We therefore train one vision–language model f_θ with role-conditioned, sample-aligned supervision. For a given image, the Policy record pairs (x, q) with a target caption; the Generator record pairs the image, reference captions, and candidate captions with a target rubric; and the Judge records pair a candidate caption and rubric items with criterion-level decisions and rationales. The records are linked by their underlying image and caption instances, while role prompts distinguish the three tasks.

Let D_{SFT} be the resulting mixed dataset. All three roles update the same parameters through the autoregressive objective

$$\mathcal{L}_{\mathrm{SFT}}(\theta) = \mathbb{E}_{(u,v)\sim D_{\mathrm{SFT}}}\left[-\sum_{\ell=1}^{|v|}\log p_\theta(v_\ell \mid u, v_{<\ell})\right], \tag{1}$$

where u comprises a role prompt and its task inputs, and v is the corresponding supervision target. We initialize both the online Policy and the momentum model from the resulting checkpoint, $\theta_0 = \phi_0 = \theta_{\mathrm{SFT}}$. During Stage II, the Generator and Judge use the same parameters ϕ under their respective role prompts; only the Policy parameters θ receive RL gradients. This initialization establishes the rubric-role capabilities before parameter transfer begins.

3.3 RUBRIC-GUIDED POLICY OPTIMIZATION WITH MOMENTUM ADAPTATION

Stage II adapts the content of each rubric to current Policy candidates and transfers Policy parameter changes gradually to the two pretrained rubric roles.

Online rubric construction. At Policy update t, we sample $N = 16$ captions from the behavior Policy, $Y_t = \{y_i\}_{i=1}^N$, and select a random subset $S_t \subseteq Y_t$ of $M = 4$ captions for rubric construction. Given the reference captions C^{ref}, the Generator produces $R_t = G_{\phi_t}(x, C^{\mathrm{ref}}, S_t) = \{(c_k, e_k, w_k)\}_{k=1}^{K_t}$, where c_k is a criterion, e_k specifies how to decide whether it is satisfied, and $w_k > 0$ is its weight. The subset S_t exposes differences among current candidates, the references supply a quality anchor beyond that subset, and the image provides the factual basis for each criterion. The resulting rubric is applied to every caption in Y_t, including captions outside S_t.

The Judge receives each caption and the image-grounded rubric, without direct access to the image, and produces a decision $b_{ik} \in \{0, 1\}$ for each item, along with a rationale. A value of one means that caption y_i satisfies item k. The criterion-level decisions yield a normalized weighted reward $b_{ik} = \left[J_{\phi_t}(y_i, R_t)\right]_k$, $r_i = \frac{\sum_{k=1}^{K_t} w_k b_{ik}}{\sum_{k=1}^{K_t} w_k} \in [0, 1]$. Using the same rubric and momentum checkpoint for the whole group makes reward differences comparable across candidates. The Judge's rationales explain the item decisions, while only the binary decisions and weights contribute to r_i.

Policy optimization and momentum transfer. We standardize the rewards within each group to obtain GRPO advantages, $A_i = \frac{r_i - \mathrm{mean}_j(r_j)}{\mathrm{std}_j(r_j)+\epsilon}$, where $\epsilon > 0$ ensures numerical stability. The Policy optimizes the GRPO clipped surrogate using these advantages (Shao et al., 2024). Writing $\mathbf{r}_t = (r_1, \ldots, r_N)$, we summarize Policy update as $\theta_{t+1} = \mathrm{GRPOUpdate}(\theta_t; x, q, Y_t, \mathrm{stopgrad}(\mathbf{r}_t))$. The Generator and Judge receive no gradients from this update. Their capabilities come from Stage I, while transferring changes in shared visual and linguistic representations may help them remain coordinated with the evolving Policy. The Generator and Judge share momentum parameters ϕ_t, initialized as $\phi_0 = \theta_0$. After every H Policy updates, the shared momentum model is updated by

$$\phi_{t+1} = \begin{cases} m\phi_t + (1-m)\theta_{t+1}, & (t+1) \bmod H = 0, \\ \phi_t, & \text{otherwise}, \end{cases} \tag{2}$$

where $m \in [0, 1]$ is the momentum coefficient.

Synchronization and adaptation. Since the RL objective optimizes caption generation rather than rubric quality, we do not assume that copying Policy parameters improves either rubric role. We use a controlled transfer, inspired by MoCo (He et al., 2020). We define the transfer rule as an exponential moving average (EMA), aiming to balance adaptation with retention of those capabilities:

$$\mathcal{S}(\phi_t, \theta_{t+1}) = m\phi_t + (1-m)\theta_{t+1}. \tag{3}$$

Here $m \in [0, 1]$ is the momentum coefficient; the two terms retain the previous momentum parameters and incorporate the updated Policy parameters, respectively.

EMA retains part of the previous rubric-role state while incorporating a fraction of the updated Policy parameters. The update occurs after the current group has been scored and before a subsequent group is evaluated. Both rubric roles therefore use the same parameter version to formulate and execute each rubric. Rubric content changes with the candidate group, whereas the model parameters change only at synchronization boundaries. The cases $m = 1$, $(m, H) = (0, 1)$, and $m = 0$ with $H > 1$ correspond to frozen rubric roles, copying after every Policy update, and periodic full copying, respectively. The tracking and smoothing effects of intermediate values are analyzed below; whether the rubric roles retain their task performance requires empirical evaluation.

3.4 Tracking–Stability Analysis

We analyze how the momentum coefficient m and synchronization interval H balance tracking of the current Policy against smooth updates of the shared rubric model. Let $\bar{\theta}_s$ and $\bar{\phi}_s$ be Policy and momentum parameters at synchronization boundary s. Define the tracking error $e_s = \|\bar{\theta}_s - \bar{\phi}_s\|$, the Policy displacement $\Delta_s = \|\bar{\theta}_s - \bar{\theta}_{s-1}\|$, and the momentum-update magnitude $u_s = \|\bar{\phi}_s - \bar{\phi}_{s-1}\|$.

Proposition 1 (Tracking–stability trade-off). Suppose $0 \le m < 1$, each Policy update satisfies $\|\theta_{t+1} - \theta_t\| \le \delta$, and synchronization occurs every H updates. Then

$$e_s \le m^s e_0 + \frac{m(1-m^s)}{1-m} H\delta, u_s \le (1-m)(e_{s-1} + \Delta_s). \tag{4}$$

Shared initialization gives $e_0 = 0$, and hence $e_s \le mH\delta/(1-m)$. Thus, the Policy–momentum parameter gap remains bounded under bounded Policy updates, while m and H jointly determine how closely the rubric roles track the current Policy.

To quantify stochastic smoothing, we consider the local approximation $\bar{\theta}_s - \bar{\theta}_{s-1} = Hv + \varepsilon_s$, $\mathbb{E}[\varepsilon_s] = 0$,$\mathrm{Cov}(\varepsilon_s) = H\Sigma$, where the mean drift v is approximately constant within the local analysis window and ε_s is independent across non-overlapping synchronization intervals. The covariance assumption models the aggregation of H approximately independent per-step perturbations with locally stationary covariance. These stochastic assumptions are used only for the smoothing analysis and are not required for Proposition 1. Let $q_s = \bar{\phi}_s - \bar{\phi}_{s-1}$. In the stationary regime, $\mathbb{E}[q_s] = Hv$, $\mathrm{Cov}(q_s) = \frac{1-m}{1+m} H\Sigma$. Compared with periodic full copying ($m = 0$), EMA therefore reduces the stochastic update covariance by the factor $(1-m)/(1+m)$, while preserving the long-run mean drift. A larger m produces smoother rubric-role updates but increases tracking lag, whereas a larger H increases the drift and noise accumulated between synchronizations. Appendix D gives the full finite-step and stationary derivations. These results describe parameter tracking and smoothing, rather than guaranteeing that caption optimization improves rubric construction or judging.

4 Experiments

4.1 Experimental Setup

Implementation Details. We use Qwen3-VL-8B-Instruct (Bai et al., 2025) as the backbone and perform full-parameter training on H100 GPUs (80GB). Both SFT and RL datasets contain equal proportions of data from PixMo-Cap (Deitke et al., 2024) and DenseFusion (Li et al., 2024). For SFT, we use 20,000 supervision records generated by GPT-5.5 (OpenAI, 2026): 10,000 for the Policy and 5,000 each for the Generator and Judge. We train for 300 steps with a learning rate of 1×10^{-5} and a global batch size of 64. For RL, we optimize the Policy with GRPO (Shao et al., 2024) on 5,000 samples for 600 steps at a learning rate of 1×10^{-6}. We use a per-device batch size

Table 1: Pairwise win rates (%) against Qwen3-VL-8B across five captioning benchmarks. Average denotes the benchmark mean. Best results are bold; second-best results are underlined.

Method	PixMo-Cap	DenseFusion	CapArena	CompreCap	DOCCI	Average
Captioning Baselines						
ShareGPT4V-7B	2.02	1.80	2.67	5.21	2.41	2.82
RICO-Flash-7B	8.89	5.41	9.00	9.87	9.04	8.44
OmniCaptioner-8B	10.91	10.62	12.17	10.05	11.24	11.00
JoyCaption-8B	11.72	4.81	12.67	25.85	14.46	13.90
MetaCaptioner-8B	22.63	23.25	19.00	32.50	18.47	23.17
CapRL-InternVL-8B	23.03	16.83	15.67	25.67	15.26	19.29
CapRL-Qwen3VL-4B	42.22	39.48	48.83	34.83	50.00	43.07
Qwen3-VL-32B	64.85	63.73	62.00	56.19	57.23	60.80
Rubric-based Methods						
RubiCap	<u>70.14</u>	<u>67.80</u>	<u>72.33</u>	**63.57**	<u>75.80</u>	<u>69.93</u>
EvoLM	59.19	53.71	56.17	48.47	62.63	56.03
Our Framework						
SFT (shared)	55.15	53.31	51.83	49.19	54.22	52.74
Ours	**73.94**	**70.34**	**78.00**	<u>63.38</u>	**78.51**	**72.83**

Table 2: Caption-based question answering across five benchmarks. Average is the mean across benchmarks. Bold and underlined values indicate the best and second-best results, respectively.

Method	CaptionQA	BLINK	TextVQA	DocVQA	ChartQA	Average
Base	80.41	47.05	<u>58.00</u>	69.18	61.28	63.18
RubiCap	<u>83.28</u>	49.05	56.85	<u>77.90</u>	**64.44**	<u>66.31</u>
Qwen3-VL-32B	82.92	**50.94**	56.69	75.06	62.16	65.55
EvoLM	82.00	49.00	57.08	77.21	62.44	65.55
SFT (shared)	81.05	47.69	56.44	76.13	63.16	64.89
Ours	**84.24**	<u>50.10</u>	**58.33**	**79.05**	<u>64.12</u>	**67.17**

of 8 and two gradient accumulation steps. Both stages use AdamW (Loshchilov & Hutter, 2019) with cosine learning rate decay and a 5% warmup ratio.

Benchmarks. We evaluate both caption quality and downstream task utility. Caption quality evaluation covers PixMo-Cap (Deitke et al., 2024), DenseFusion (Li et al., 2024), CapArena (Cheng et al., 2025), CompreCap (Lu et al., 2025a), and DOCCI (Onoe et al., 2024), with blind ranking and ablation studies conducted on PixMo-Cap and DenseFusion. Downstream evaluation spans caption utility (CaptionQA (Yang et al., 2026b)), visual perception and reasoning (BLINK (Fu et al., 2024)), and scene-text understanding (TextVQA (Singh et al., 2019)). It also includes document understanding (DocVQA (Mathew et al., 2021)) and chart understanding and reasoning (ChartQA (Masry et al., 2022)). These complementary tasks test whether generated captions preserve visual information needed to answer diverse questions, complementing direct assessments of caption quality.

Baselines. We compare captioning baselines with rubric-based methods. Captioning baselines include ShareGPT4V-7B (Chen et al., 2024), RICO-Flash-7B (Wang et al., 2025), OmniCaptioner-8B (Lu et al., 2025b), JoyCaption-8B (fpgaminer), MetaCaptioner-8B (Lei et al., 2026), the InternVL-8B and Qwen3-VL-4B variants of CapRL (Xing et al., 2026), and the larger Qwen3-VL-32B-Instruct (Bai et al., 2025). For rubric-based methods, we adapt and reproduce RubiCap (Huang et al., 2026) and EvoLM (Li et al., 2026) for our captioning task and training data. We select EvoLM as the alternating-training baseline because the original method shares Policy and Generator parameters, offering a closer comparison to our shared initialization. DynamicRubric (Wang et al., 2026) also follows an alternating-training approach, but we do not reproduce it because its Generator training requires additional ranking annotations for anchor responses. Our RubiCap reproduction uses Gemini 3.1 Pro as the Rubric Generator and GPT-5.6-terra as the Rubric Judge. We report the shared SFT model to assess gains from subsequent reinforcement learning.

Evaluation Metrics. Following CapArena (Cheng et al., 2025), an LLM judge compares each method's captions against Qwen3-VL-8B-Instruct; we report win rates. Following RubiCap (Huang et al., 2026), we blind-rank captions from five representative methods and report rank distributions and quality scores. For question answering, we follow Prism (Qiao et al., 2024), as adopted in

Table 3: Component ablation on PixMo-Cap and DenseFusion, reported as win rates (%). Bold and underlined values indicate the best and second-best available results, respectively.

ID	SFT initialization	Rubrics	Rubric-role update	PixMo-Cap	DenseFusion	Average
M1	Policy only	—	—	57.37	56.91	57.14
M2	Shared across roles	—	—	55.15	53.31	54.23
M3	Separate per role	Offline, fixed	No momentum update	70.74	66.00	68.37
M4	Separate per role	Online	No momentum update	71.31	<u>67.13</u>	<u>69.22</u>
M5	Separate per role	Online	EMA ($m = 0.99$, $H = 1$)	65.06	61.80	63.43
M6	Shared across roles	Online	No momentum update	68.48	64.73	66.61
M7	Shared across roles	Online	Full synchronization	<u>71.72</u>	65.73	68.73
M8	Shared across roles	Online	EMA ($m = 0.99$, $H = 1$)	**73.94**	**70.34**	**72.14**

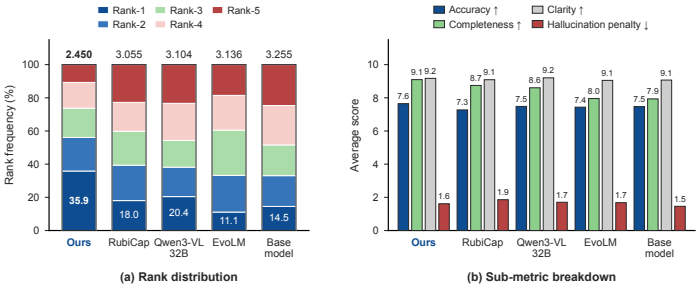


Figure 3: Blind ranking on PixMo-Cap and DenseFusion: rank distributions (left) and caption quality scores (right), averaged across the two datasets.

Table 4: Role transfer; EMA $(m, H) = (0.99, 1)$.

Generator	Judge	PixMo Cap	Dense Fusion	Average
Frozen	Frozen	68.48	64.73	66.61
EMA	Frozen	71.14	<u>67.00</u>	<u>69.07</u>
Frozen	EMA	68.27	63.40	65.84
Full sync	Full sync	<u>71.72</u>	65.73	68.73
EMA	EMA	**73.94**	**70.34**	**72.14**

CapRL (Xing et al., 2026), using a fixed text-only model with captions as its sole source of visual information. We report benchmark scores; evaluation prompts and the question-answering protocol are provided in Appendices A.1 and A.3.

4.2 Main Results

Pairwise Caption Comparison. Table 1 shows that MoCo Rubric achieves a 72.83% average win rate across five benchmarks, exceeding Qwen3-VL-32B by 12.03 percentage points with an 8B backbone. RubiCap and EvoLM also surpass the specialized captioners on average, while MoCo Rubric further exceeds them by 2.90 and 16.80 points, respectively, despite RubiCap's use of proprietary models for rubric generation and judging. EvoLM trails shared SFT on CompreCap (48.47% versus 49.19%), illustrating that dynamic rubric training does not improve every benchmark. The advantage persists with GPT-5.6 Sol as the evaluator, and both LLM evaluators show high agreement with human preferences on the tested caption pairs (Appendices B.1 and B.2).

Caption-Based Visual Question Answering. Using captions as the sole visual input to a fixed text-only model, MoCo Rubric achieves the highest average score of 67.17 in Table 2, exceeding RubiCap and shared SFT by 0.86 and 2.28 points. It ranks first on CaptionQA, TextVQA, and DocVQA and second on BLINK and ChartQA, showing that the caption gains also benefit downstream question answering.

Blind Ranking. In the joint evaluation of five anonymized methods, MoCo Rubric achieves the best mean rank (2.450) and highest first-place share (35.9%) in Figure 3. It leads in accuracy and completeness, has clarity comparable to Qwen3-VL-32B, and improves on RubiCap's mean rank (3.055) and hallucination penalty (1.859 versus 1.614).

4.3 Analysis

Which Components Drive the Gains? We ablate components and combinations to assess contributions to caption quality. In Table 3, replacing fixed offline rubrics with online rubrics (M3→M4) improves average win rate by 0.85 pp, supporting criteria adapted to current Policy candidates. Shared initialization alone offers no advantage: shared SFT underperforms Policy-only SFT (M2

vs. M1), and remains below separate initialization when rubric roles are frozen (M6 vs. M4). Its benefit emerges with momentum transfer. Under separate initialization, EMA reduces average win rate from 69.22% to 63.43% (M4→M5); with shared initialization, it raises win rate from 66.61% to 72.14% (M6→M8). This contrast supports a common parameter foundation for transferring Policy updates to both rubric roles. EMA outperforms full synchronization by 3.41 pp (M7→M8), indicating that gradual transfer contributes to the gains. These results highlight complementary contributions: online rubrics adapt evaluation criteria to current candidates, while shared initialization and momentum updates jointly support cross-role parameter transfer. Training diagnostics further show fewer rollout groups with zero within-group reward variation under our method (Appendix B.3).

Which Roles Benefit from Momentum Transfer? Starting from the shared SFT checkpoint, we apply EMA to either rubric role or both. Table 4 shows that Generator-only EMA raises the average win rate from 66.61% to 69.07%, whereas updating only the Judge lowers it to 65.84%. Joint EMA reaches 72.14%, exceeding Generator-only EMA and full synchronization (68.73%), suggesting that Judge updates help when coordinated with Generator updates through momentum transfer.

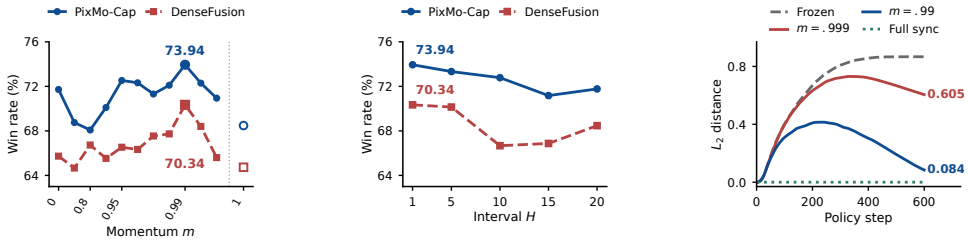


Figure 4: **Left:** Momentum coefficient analysis at $H = 1$, with tested coefficients equally spaced and the frozen reference ($m = 1$) shown separately. **Middle:** Synchronization interval analysis at $m = 0.99$. Markers indicate measured results; lines connect adjacent tested settings. **Right:**Policy–momentum tracking at $H = 1$. Frozen and full-sync curves are boundary references.

How Do Transfer Magnitude and Frequency Affect Performance?

The momentum coefficient m controls transfer magnitude, and the interval H controls transfer frequency. Figure 4 shows that $m = 0.99$ performs best with stepwise synchronization; at this coefficient, $H = 5$ remains close to $H = 1$. Table 5 reports the highest mean win rate of 72.14% at $(m, H) = (0.99, 1)$, with 71.79% at $(0.99, 5)$; longer intervals reduce win rates at $m = 0.99$, although the effect of H is not monotonic at other coefficients. These results motivate selecting m and H jointly and are consistent with the tracking–stability trade-off in Section 3.4, without establishing it as the cause of the performance differences.

Table 5: Equal-weight mean win rate (%) across PixMo-Cap and DenseFusion in the joint $m \times H$ sweep.

m	$H = 1$	$H = 5$	$H = 20$
0.95	69.58	68.18	69.29
0.99	**72.14**	71.79	70.18
0.999	68.28	66.67	69.48

How Closely Does the Momentum Model Track the Policy? Figure 4 shows the parameter distance between the Policy and the momentum model at H = 1. In separate runs with m = 0.99 and m = 0.999, the distance initially increases and then levels off or declines, reaching 0.084 and 0.605, respectively, at step 600. At that step, the frozen reference computed from the m = 0.99 Policy trajectory has a distance of 0.867, while the ideal reference with full synchronization at every step has zero distance. Relative to the frozen reference on the same Policy trajectory, EMA reduces the parameter distance in later training. This observation provides empirical support for the tracking error analysis in Proposition 1 of Section 3.4.

5 Conclusion

We presented a framework that coordinates caption generation, rubric construction, and judging to address cross-role mismatch in detailed image captioning. Shared multi-task SFT establishes a foundation, while online rubrics and EMA parameter updates sustain adaptation as the Policy evolves. Experiments show stronger overall performance than existing methods across complementary evaluations, with gains in caption quality and downstream utility. These results support coordinated rubric adaptation as an approach to reinforcement learning for detailed image captioning.

AI USE STATEMENT

AI systems, including LLMs, did not contribute to the development of the research ideas or the writing of this paper to a degree that would warrant authorship or contributor status. In this work, LLMs were used for open-ended task evaluation and served as objects of study.

ETHICS STATEMENT

We study image captioning using publicly available datasets, and our work does not involve private or sensitive data. The proposed training procedure is designed as a general-purpose optimization method. Within the scope of this study, we have not identified specific ethical risks related to fairness, bias, discrimination, privacy, or security. This research was conducted in accordance with established standards of research integrity.

REPRODUCIBILITY STATEMENT

We provide training details, implementation settings for our method, and evaluation procedures in the main text and appendix to support reproduction of the reported results.

REFERENCES

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

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

David M. Chan, Suzanne Petryk, Joseph E. Gonzalez, Trevor Darrell, and John Canny. CLAIR: Evaluating image captions with large language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 13638–13646, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.841. URL https://aclanthology.org/2023.emnlp-main.841/.

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

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.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, Jiasen Lu, Taira Anderson, Erin Bransom, Kiana Ehsani, Huong Ngo, YenSung Chen, Ajay Patel, Mark Yatskar, Chris Callison-Burch, Andrew Head, Rose Hendrix, Favyen Bastani, Eli VanderBilt, Nathan Lambert, Yvonne Chou, Arnavi Chheda, Jenna Sparks, Sam Skjonsberg, Michael Schmitz, Aaron Sarnat, Byron Bischoff, Pete Walsh, Chris Newell, Piper Wolters, Tanmay Gupta, Kuo-Hao Zeng, Jon Borchardt, Dirk Groeneveld, Crystal Nam, Sophie Lebrecht, Caitlin Wittlif, Carissa Schoenick, Oscar Michel, Ranjay Krishna, Luca Weihs, Noah A. Smith, Hannaneh Hajishirzi, Ross Girshick, Ali Farhadi, and Aniruddha Kembhavi. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models, 2024. URL https://arxiv.org/abs/2409.17146.

Hongxin Ding, Baixiang Huang, Yue Fang, Weibin Liao, Zheng Li, Jinyang Zhang, Zhijing Wu, Junfeng Zhao, and Yasha Wang. Evorubrics: Dynamic rubrics as rewards via adversarial co-evolution for llm reinforcement learning, 2026. URL https://arxiv.org/abs/2606.23038.

Hongyuan Dong, Jiawen Li, Bohong Wu, Jiacong Wang, Yuan Zhang, and Haoyuan Guo. Benchmarking and improving detail image caption, 2024. URL https://arxiv.org/abs/2405.19092.

fpgaminer. JoyCaption. GitHub repository. URL https://github.com/fpgaminer/joycaption. Accessed September 26, 2026.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A. Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive, 2024. URL https://arxiv.org/abs/2404.12390.

Xin Guan, Xiaomeng Hu, Shen Huang, Zhenyi Wang, Bo Zhang, Zijian Li, Pengjun Xie, Bo Liu, and Jiuxin Cao. Evorubric: Self-evolving rubric-driven rl for open-ended generation, 2026. URL https://arxiv.org/abs/2605.29847.

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.

Danna Gurari, Yinan Zhao, Meng Zhang, and Nilavra Bhattacharya. Captioning images taken by people who are blind. In *European Conference on Computer Vision*, pp. 417–434. Springer, 2020.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning, 2020. URL https://arxiv.org/abs/1911.05722.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 7514–7528, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.595. URL https://aclanthology.org/2021.emnlp-main.595/.

Tzu-Heng Huang, Sirajul Salekin, Javier Movellan, Frederic Sala, and Manjot Bilkhu. Rubicap: Rubric-guided reinforcement learning for dense image captioning. *arXiv preprint arXiv:2603.09160*, 2026.

Zenan Huang, Yihong Zhuang, Guoshan Lu, Zeyu Qin, Haokai Xu, Tianyu Zhao, Ru Peng, Jiaqi Hu, Zhanming Shen, Xiaomeng Hu, Xijun Gu, Peiyi Tu, Jiaxin Liu, Wenyu Chen, Yuzhuo Fu, Zhiting Fan, Yanmei Gu, Yuanyuan Wang, Zhengkai Yang, Jianguo Li, and Junbo Zhao. Reinforcement learning with rubric anchors, 2025. URL https://arxiv.org/abs/2508.12790.

Ruipeng Jia, Yunyi Yang, Wen Wang, Yuxin Wu, Yongbo Gai, Siyuan Tao, Mengyu Zhou, Jianhe Lin, Xiaoxi Jiang, and Guanjun Jiang. Open rubric system: Scaling reinforcement learning with pairwise adaptive rubric, 2026. URL https://arxiv.org/abs/2602.14069.

Yebin Lee, Imseong Park, and Myungjoo Kang. FLEUR: An explainable reference-free evaluation metric for image captioning using a large multimodal model. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3732–3746, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.205. URL https://aclanthology.org/2024.acl-long.205/.

Zhenxin Lei, Zhangwei Gao, Changyao Tian, Erfei Cui, Guanzhou Chen, Danni Yang, Yuchen Duan, Zhaokai Wang, Wenhao Li, Weiyun Wang, et al. Metacaptioner: Towards generalist visual captioning with open-source suites. In *International Conference on Learning Representations*, volume 2026, pp. 102906–102945, 2026.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. URL https://arxiv.org/abs/2201.12086.

Shuyue Stella Li, Rui Xin, Teng Xiao, Yike Wang, Rulin Shao, Zoey Hao, Melanie Sclar, Sewoong Oh, Faeze Brahman, Pang Wei Koh, and Yulia Tsvetkov. Evolm: Self-evolving language models through co-evolved discriminative rubrics, 2026. URL https://arxiv.org/abs/2605.03871.

Xiaotong Li, Fan Zhang, Haiwen Diao, Yueze Wang, Xinlong Wang, and Ling-Yu Duan. Densefusion-1m: Merging vision experts for comprehensive multimodal perception, 2024. URL https://arxiv.org/abs/2407.08303.

Tianci Liu, Ran Xu, Tony Yu, Ilgee Hong, Carl Yang, Tuo Zhao, and Haoyu Wang. Openrubrics: Towards scalable synthetic rubric generation for reward modeling and llm alignment, 2026. URL https://arxiv.org/abs/2510.07743.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019. URL https://arxiv.org/abs/1711.05101.

Fan Lu, Wei Wu, Kecheng Zheng, Shuailei Ma, Biao Gong, Jiawei Liu, Wei Zhai, Yang Cao, Yujun Shen, and Zheng-Jun Zha. Benchmarking large vision-language models via directed scene graph for comprehensive image captioning, 2025a. URL https://arxiv.org/abs/2412.08614.

Yiting Lu, Jiakang Yuan, Zhen Li, Shitian Zhao, Qi Qin, Xinyue Li, Le Zhuo, Licheng Wen, Dongyang Liu, Yuewen Cao, Xiangchao Yan, Xin Li, Tianshuo Peng, Shufei Zhang, Botian Shi, Tao Chen, Zhibo Chen, Lei Bai, Peng Gao, and Bo Zhang. Omnicaptioner: One captioner to rule them all, 2025b. URL https://arxiv.org/abs/2504.07089.

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

Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for vqa on document images, 2021. URL https://arxiv.org/abs/2007.00398.

Yasumasa Onoe, Sunayana Rane, Zachary Berger, Yonatan Bitton, Jaemin Cho, Roopal Garg, Alexander Ku, Zarana Parekh, Jordi Pont-Tuset, Garrett Tanzer, Su Wang, and Jason Baldridge. Docci: Descriptions of connected and contrasting images, 2024. URL https://arxiv.org/abs/2404.19753.

OpenAI. Introducing GPT-5.5, 2026. URL https://openai.com/index/introducing-gpt-5-5/. Published April 23, 2026.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin (eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

Yuxuan Qiao, Haodong Duan, Xinyu Fang, Junming Yang, Lin Chen, Songyang Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. Prism: A framework for decoupling and assessing the capabilities of vlms. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 111863–111898. Curran Associates, Inc., 2024. doi: 10.52202/079017-3552. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/cac9e747a1d480c78312226959566cef-Paper-Conference.pdf.

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

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii (eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4035–4045, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1437. URL https://aclanthology.org/D18-1437/.

Rulin Shao, Akari Asai, Shannon Zejiang Shen, Hamish Ivison, Varsha Kishore, Jingming Zhuo, Xinran Zhao, Molly Park, Samuel G. Finlayson, David Sontag, Tyler Murray, Sewon Min, Pradeep Dasigi, Luca Soldaini, Faeze Brahman, Wen tau Yih, Tongshuang Wu, Luke Zettlemoyer, Yoon Kim, Hannaneh Hajishirzi, and Pang Wei Koh. Dr tulu: Reinforcement learning with evolving rubrics for deep research, 2026. URL https://arxiv.org/abs/2511.19399.

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.

William F. Shen, Xinchi Qiu, Chenxi Whitehouse, Lisa Alazraki, Shashwat Goel, Francesco Barbieri, Timon Willi, Akhil Mathur, and Ilias Leontiadis. Rethinking rubric generation for improving llm judge and reward modeling for open-ended tasks, 2026. URL https://arxiv.org/abs/2602.05125.

Leheng Sheng, Wenchang Ma, Ruixin Hong, Xiang Wang, An Zhang, and Tat-Seng Chua. Reinforcing chain-of-thought reasoning with self-evolving rubrics, 2026. URL https://arxiv.org/abs/2602.10885.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read, 2019. URL https://arxiv.org/abs/1904.08920.

Vasu Singla, Kaiyu Yue, Sukriti Paul, Reza Shirkavand, Mayuka Jayawardhana, Alireza Ganjdanesh, Heng Huang, Abhinav Bhatele, Gowthami Somepalli, and Tom Goldstein. From pixels to prose: A large dataset of dense image captions, 2024. URL https://arxiv.org/abs/2406.10328.

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.

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

Vijay Viswanathan, Yanchao Sun, Shuang Ma, Xiang Kong, Meng Cao, Graham Neubig, and Tongshuang Wu. Checklists are better than reward models for aligning language models, 2025. URL https://arxiv.org/abs/2507.18624.

Beining Wang, Weihang Su, Hongtao Tian, Hao Kong, Tao Yang, Ting Yao, Qingyi Pan, Yueyue Wu, Qingyao Ai, Min Zhang, and Yiqun Liu. Co-evolving llm evaluators and policies via dynamicrubric, 2026. URL https://arxiv.org/abs/2607.20083.

Yuchi Wang, Yishuo Cai, Shuhuai Ren, Sihan Yang, Linli Yao, Yuanxin Liu, Yuanxing Zhang, Pengfei Wan, and Xu Sun. Rico: Improving accuracy and completeness in image recaptioning via visual reconstruction, 2025. URL https://arxiv.org/abs/2505.22613.

Long Xing, Qidong Huang, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Jinsong Li, Shuangrui Ding, Weiming Zhang, Nenghai Yu, et al. Scalecap: Inference-time scalable image captioning via dual-modality debiasing. *arXiv preprint arXiv:2506.19848*, 2025.

Long Xing, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Jianze Liang, Qidong Huang, Jiaqi Wang, Feng Wu, and Dahua Lin. Caprl: Stimulating dense image caption capabilities via reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 13066–13093, 2026.

Ran Xu, Tianci Liu, Zihan Dong, Tony Yu, Ilgee Hong, Carl Yang, Linjun Zhang, Tao Zhao, and Haoyu Wang. Alternating reinforcement learning for rubric-based reward modeling in non-verifiable llm post-training, 2026. URL https://arxiv.org/abs/2602.01511.

Penghui Yang, Long Xing, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Yibin Wang, Yujie Zhou, Jiazi Bu, Jianze Liang, Qidong Huang, Jiaqi Wang, Feng Wu, and Dahua Lin. Caprl++: Unified reinforcement learning with verifiable rewards for dense image and video captioning, 2026a. URL https://arxiv.org/abs/2606.09393.

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?, 2026b. URL https://arxiv.org/abs/2511.21025.

Shaokai Ye, Vasileios Saveris, Yihao Qian, Jiaming Hu, Elmira Amirloo, and Peter Grasch. Balcaprl: A balanced framework for rl-based mllm image captioning, 2026. URL https://arxiv.org/abs/2605.07394.

Junkai Zhang, Zihao Wang, Lin Gui, Swarnashree Mysore Sathyendra, Jaehwan Jeong, Victor Veitch, Wei Wang, Yunzhong He, Bing Liu, and Lifeng Jin. Chasing the tail: Effective rubric-based reward modeling for large language model post-training, 2026. URL https://arxiv.org/abs/2509.21500.

A Implementation and Evaluation Details

A.1 Prompt Templates

We provide the prompts for the Caption Policy, Rubric Generator, Rubric Judge, and two external caption evaluations. Per-example inputs are represented by placeholders. The Policy and Generator receive the image, denoted by `<image>` below. Both external evaluators also receive the image, whereas the Rubric Judge receives only the caption and rubrics.

A.1.1 Caption Generation

The Caption Policy generates a detailed image description using the following prompt.

System prompt

```
You are an image captioning assistant. You output only plain-text
image descriptions without any markdown formatting, headers, bullet
points, or structural elements. Your descriptions are detailed,
flowing paragraphs.
```

User prompt

```
<image>Describe this image in detail.
```

A.1.2 Rubric Generation

The Rubric Generator receives the image, two reference captions, and four sampled current-Policy captions. The following template specifies the criterion, binary evaluation rule, and weight of each rubric item.

User prompt

```
<image>
You generate strict, image-grounded rubrics for distinguishing the
quality of image captions.

You receive one image, two strong reference captions, and four
captions sampled from the current policy. The image is the final
authority. The references help identify reliable details. The policy
captions reveal quality differences that the rubrics should capture;
none of the captions is automatically correct.

Reference caption 1:
<reference_caption_1>

Reference caption 2:
<reference_caption_2>

Current-policy caption 1:
<policy_caption_1>

Current-policy caption 2:
<policy_caption_2>

Current-policy caption 3:
<policy_caption_3>

Current-policy caption 4:
<policy_caption_4>

Goal:
```

User prompt (continued)

````
Generate binary rubric items that capture important, image-verifiable
differences among the policy captions. Prioritize decision boundaries
that separate more accurate and precise captions from vague,
incomplete, incorrect, or unsupported ones. Do not optimize for
exhaustive coverage or generate easy items merely because they
describe true image content.

Instructions:
1. Inspect the image and use the references to identify salient,
reliable facts.
2. Compare the policy captions to identify any meaningful,
image-verifiable difference that affects caption quality. Consider
factual correctness, specificity, important omissions, unsupported
claims, and any other distinction that materially separates better
captions from worse ones.
3. Verify each difference against the image, then express the correct
distinction as a self-contained rubric that also applies to future
captions.
4. Preserve the most specific important fact supported by the image.
Do not weaken an exact count, visible text, identity, attribute,
action, or spatial relation into a broader condition simply to make
more captions pass.
5. Define each Pass/Fail boundary at the highest reliable specificity
supported by the image. When an important fact is precisely verifiable
and that precision matters to caption quality, vague wording, an
incorrect value, or omission must fail. Do not enforce uncertain,
unimportant, or incidental details.
6. Choose a rubric set that reflects overall caption quality: preserve
important distinctions while giving appropriate credit for accurate
central content. Avoid relying entirely on error-specific checks or
broad, easy-to-pass facts, and consolidate errors that arise from the
same underlying fact.
7. Keep each item atomic and non-redundant. Do not create overlapping
or nested criteria for the same underlying fact.
8. Judge semantic meaning rather than exact wording, except when exact
visible text, numbers, names, or counts are the fact being tested.

Illustrative example:
If the image clearly shows exactly five masks and the count materially
distinguishes caption quality, use one rubric requiring exactly five
masks. "Multiple masks", a wrong count, or omission of the important
count must fail; do not add separate broader criteria for the presence
of masks. Apply this precision-preserving principle to any reliable
distinction, not only counts.

Weights:
- 3.0 for a central fact or severe factual error.
- 2.0 for an important quality distinction.
- 1.0 for a useful secondary distinction.

Return only one valid JSON object inside one ```json code fence:
```json
{
  "rubrics": [
    {
      "criterion": "...",
      "evaluation_rule": "Pass if ...; fail if ...",
      "weight": 1.0
    }
  ]
}
```
````

User prompt (continued)

```
Do not mention the references or policy captions in the output. Do not
add extra fields or text outside the JSON code fence.
```

A.1.3 RUBRIC-BASED SCORING

The text-only Rubric Judge evaluates each caption against the supplied rubrics, returning a reason followed by a binary decision for each item. The variables in braces are filled for each request.

System prompt

```
You are a strict reward evaluator for image captions. Judge only
whether the candidate caption satisfies the supplied rubrics. The
rubrics are the ground truth. Do not reward unsupported claims. Do not
penalize issues that are not covered by the rubrics. Give one concise,
caption-grounded reason before each binary decision. Return valid JSON
only.
```

User prompt template

```
You are an image-caption reward judge.
Your task is to evaluate whether a generated caption satisfies each
supplied rubric criterion.
Judge by semantic meaning and intent, not exact wording or keyword
matching.

Generated caption:
{generated_caption}

Rubrics:
{rubrics}

Evaluation rules:
1. Evaluate every rubric independently, in the same order as listed.
2. Before pass, write a concise, specific reason grounded in the
generated caption and the current rubric. For pass=1, identify how all
required conditions are satisfied; for pass=0, identify the decisive
unmet, contradicted, or unsupported condition.
3. Set pass=1 only if the caption satisfies both the criterion and its
evaluation_rule.
4. Set pass=0 if the caption omits required information, contradicts
the rule, or contains unsupported content explicitly covered by the
rubric.
5. Judge only the requirements in the current rubric. Do not penalize
issues that the rubric does not cover.
6. Accept synonyms, different sentence structures, and other
semantically equivalent expressions.
7. Do not penalize style, fluency, grammar, or length unless the
rubric explicitly requires it.

Return exactly one valid JSON object in this format:
{example_json}
The items array must contain exactly {item_count} elements, one per
rubric, in the same order.
Each item must contain exactly two fields in this order: reason, then
pass.
Each reason must be a non-empty JSON string containing a concise,
specific, evidence-based explanation.
Each pass value must be the integer 0 or 1, not true/false and not a
string.
```

User prompt template (continued)

```
Return valid JSON only. Do not include markdown, code fences,
comments, scores, or text outside the JSON object.
```

Here {generated_caption} is the candidate caption and {item_count} is the number of rubric items, N. The {rubrics} field lists all N items in the following format; the ellipsis denotes repeated entries.

Rubric serialization

```
1. weight=<weight_1>
criterion: <criterion_1>
evaluation_rule: <evaluation_rule_1>

...

N. weight=<weight_N>
criterion: <criterion_N>
evaluation_rule: <evaluation_rule_N>
```

The {example_json} field contains exactly N example items, alternating the two reason–decision patterns below. This two-item illustration specifies the output format; it is not a judgment of a particular caption.

Output-format example for two rubric items

```
{
  "items": [
    {
      "reason": "The caption states the required subject and
      attribute.",
      "pass": 1
    },
    {
      "reason": "The caption omits a required detail.",
      "pass": 0
    }
  ]
}
```

A.1.4 Pairwise Evaluation for Win Rate

An external evaluator receives the image and two captions and returns A, B, or Tie, together with a reason. In the stored pairwise evaluation requests, Caption A is the evaluated method's output and Caption B is the Qwen3-VL-8B-Instruct baseline output.

Evaluation prompt

```
Given an image and two candidate captions, determine which caption is
better.

Evaluation guidelines:
1. Precision:
   Caption should accurately match the image.
   Penalize wrong color, quantity, spatial relation, posture, etc.

2. Informativeness:
   Caption should include salient information.
   More specific descriptions are preferred when accurate.

3. Hallucination:
```

Evaluation prompt (continued)

```
   Penalize descriptions of objects or elements absent from the image.

4. Attention to detail:
   Carefully inspect image details.

5. Assistive description:
   Imagine describing the image to a visually impaired person.

6. Reverse thinking:
   Ask what image the caption makes you imagine and whether it matches
   the actual image.

7. Ties are acceptable:
   If both captions are similarly good, output Tie.

Ignore:
- writing style or phrasing
- caption length
- grammatical variations

Caption A:
<caption_A>

Caption B:
<caption_B>

Example output:
{
  "reason": "...",
  "judgment": "A" | "B" | "Tie"
}

Return only one JSON object in exactly the same format as the example
output. Do not include markdown, code fences, or any text before or
after the JSON.
```

A.1.5 BLIND RANKING

The evaluator receives the image and five captions labeled `Caption A` through `Caption E`, without model names. The prompt below includes the four scoring dimensions, score aggregation rule, and output template used for both PixMo-Cap and DenseFusion.

Evaluation prompt

```
You are an expert image captioning evaluator. Given the image above
and the 5 captions below, rigorously assess each caption.

## Scoring
Score each caption on four dimensions (integers 0-10):

1. accuracy - Are the described objects, actions, text, colors, and
spatial relationships factually correct for this image? Penalize for
any wrong attribute, misidentified object, or incorrect action.
2. completeness - Does the caption cover all visually significant
elements (main subjects, notable actions, background context,
on-screen text if present)? Penalize for missing key details.
3. clarity - Is the caption well-written, specific, grammatically
correct, and unambiguous? Penalize for vague language or redundancy.
```

Evaluation prompt (continued)

```
4. hallucination_penalty - Does the caption assert things NOT visible
in the image? 0 = zero hallucination; 10 = pervasive fabrication. Be
strict: even plausible but unverifiable claims count as mild
hallucination (2-4). This score is applied as a penalty.

Compute: total_score = (accuracy + completeness + clarity) / 3.0 -
hallucination_penalty x 1.5

## Captions to Evaluate
Caption A:
<caption_A>

Caption B:
<caption_B>

Caption C:
<caption_C>

Caption D:
<caption_D>

Caption E:
<caption_E>

## Output Format
Respond ONLY with a single valid JSON object - no markdown fences, no
extra text.
{
  "assessments": {
    "Caption A": {
      "justification": "<2-3 sentences citing specific visual evidence
      from the image>",
      "accuracy": "<int 0-10>",
      "completeness": "<int 0-10>",
      "clarity": "<int 0-10>",
      "hallucination_penalty": "<int 0-10>",
      "total_score": "<float>"
    },
    "Caption B": {
      "justification": "<2-3 sentences citing specific visual evidence
      from the image>",
      "accuracy": "<int 0-10>",
      "completeness": "<int 0-10>",
      "clarity": "<int 0-10>",
      "hallucination_penalty": "<int 0-10>",
      "total_score": "<float>"
    },
    "Caption C": {
      "justification": "<2-3 sentences citing specific visual evidence
      from the image>",
      "accuracy": "<int 0-10>",
      "completeness": "<int 0-10>",
      "clarity": "<int 0-10>",
      "hallucination_penalty": "<int 0-10>",
      "total_score": "<float>"
    },
    "Caption D": {
      "justification": "<2-3 sentences citing specific visual evidence
      from the image>",
      "accuracy": "<int 0-10>",
      "completeness": "<int 0-10>",
      "clarity": "<int 0-10>",
      "hallucination_penalty": "<int 0-10>",
```

Evaluation prompt (continued)

```
      "total_score": "<float>"
    },
    "Caption E": {
      "justification": "<2-3 sentences citing specific visual evidence
      from the image>",
      "accuracy": "<int 0-10>",
      "completeness": "<int 0-10>",
      "clarity": "<int 0-10>",
      "hallucination_penalty": "<int 0-10>",
      "total_score": "<float>"
    }
  },
  "ranking": [
    "Caption A",
    "Caption B",
    "Caption C",
    "Caption D",
    "Caption E"
  ]
}

The "ranking" list must contain all 5 caption labels ordered from best
(index 0) to worst (index 4), sorted strictly by total_score
descending.
```

A.2 Reproduction of RubiCap and EvoLM

A.2.1 RubiCap

We adapt RubiCap (Huang et al., 2026) to our captioning setting and initialize the Caption Policy from the same shared-SFT checkpoint used by our method. Before RL training, Gemini 3.1 Pro constructs image-specific rubrics using reference captions generated by Gemini 3.1 Pro and GPT-5.6-terra. These rubrics are generated once and reused throughout training. For each sampled caption, GPT-5.6-terra serves as the Rubric Judge, checking whether it satisfies each criterion and aggregating the weighted binary judgments into a scalar reward. The Policy is then optimized with GRPO, while the Judge remains fixed and the rubrics are neither regenerated nor updated. This preserves RubiCap's offline rubric-guided optimization while adapting its initialization and evaluation models to our experimental setting.

A.2.2 EvoLM

We adapt EvoLM's (Li et al., 2026) alternating Policy–Generator optimization and self-constructed preference pairs to our captioning setting. The Policy, Generator, and Judge are initialized from the same shared-SFT checkpoint, with the Judge kept fixed throughout training. Rubric generation uses the same prompt and inputs as our method. We alternate between 50 steps of Policy optimization and 50 steps of Generator optimization. For Generator training, preference pairs consist of captions from the current Policy and a checkpoint 50 Policy-update steps earlier, treating the current captions as preferred. Following EvoLM, the Generator reward combines the Judge-score difference between preferred and dispreferred captions with a format-validity reward, weighted by 0.7 and 0.3, respectively. Both optimization phases use GRPO, with 600 total updates across the two roles.

A.3 Caption-Based Question Answering Protocol

We follow the decoupled VQA protocol of Prism (Qiao et al., 2024), as adopted in CapRL (Xing et al., 2026), to evaluate the utility of generated captions for downstream question answering. Each captioning model first describes the input image without access to the associated questions, using the caption generation prompt in Appendix A.1. A fixed text-only language model then receives the generated description and a question, together with answer options when applicable. The description serves as its sole source of visual information.

We evaluate this pipeline on CaptionQA, BLINK, TextVQA, DocVQA, and ChartQA. The answering model is held fixed across captioning methods, and its predictions are scored against the reference answers using the corresponding evaluation procedure for each benchmark. This setting measures how effectively captions preserve information needed for downstream questions under a common answering model.

B Supplementary Evaluation Results

B.1 Evaluation with GPT-5.6

We repeat the pairwise caption comparison in Table 1 with GPT-5.6 Sol (reasoning effort: none) as the evaluator in place of Gemini 3.1 Pro. We keep the generated captions, image–caption pairs, comparison prompt, and Qwen3-VL-8B-Instruct baseline fixed, and use the same win-rate metric. In every comparison, Caption A is the evaluated method's output and Caption B is the Qwen3-VL-8B-Instruct baseline output. Table 6 reports the results.

Table 6: Pairwise win rates (%) against Qwen3-VL-8B with GPT-5.6 Sol as the evaluator. Average denotes the benchmark mean. Best results are bold; second-best results are underlined.

Method	PixMo-Cap	DenseFusion	CapArena	CompreCap	DOCCI	Average
Captioning Baselines						
ShareGPT4V-7B	1.00	0.20	0.50	1.25	1.00	0.79
RICO-Flash-7B	5.02	4.02	5.34	8.06	4.20	5.33
OmniCaptioner-8B	5.62	3.41	6.34	7.89	9.80	6.61
JoyCaption-8B	11.65	6.43	11.52	31.18	13.60	14.87
MetaCaptioner-8B	14.06	10.64	13.52	20.97	14.80	14.80
CapRL-InternVL-8B	19.88	18.67	14.52	23.12	14.80	18.20
CapRL-Qwen3VL-4B	29.12	31.33	24.04	28.32	28.80	28.32
Qwen3-VL-32B	52.61	50.60	49.58	52.33	48.20	50.67
Rubric-based Methods						
RubiCap	<u>60.44</u>	<u>60.64</u>	<u>62.77</u>	<u>62.37</u>	<u>64.60</u>	<u>62.16</u>
EvoLM	59.84	54.82	59.27	58.06	60.40	58.48
Our Framework						
SFT (shared)	56.83	55.42	57.76	54.84	60.20	57.01
Ours	**67.47**	**67.67**	**69.45**	**69.18**	**70.20**	**68.79**

MoCo Rubric ranks first on all five benchmarks under GPT-5.6 Sol, with an average win rate of 68.79%. It exceeds RubiCap by 6.63 percentage points, EvoLM by 10.31 points, and Qwen3-VL-32B by 18.12 points on average. RubiCap ranks second on each benchmark, while both RubiCap and EvoLM outperform the strongest captioning baseline, Qwen3-VL-32B, on average.

Although the absolute win rates vary with the evaluator, MoCo Rubric has the highest average under both judges (72.83% with Gemini 3.1 Pro and 68.79% with GPT-5.6 Sol). This consistency supports the compatibility of the pairwise evaluation protocol with different LLM evaluators and shows that our method's relative advantage persists across the two judges. Each benchmark result uses images with valid judgments for all 12 methods (498, 498, 599, 558, and 500 images, respectively); 67 of 31,920 attempted comparisons were excluded by the evaluator's content filter.

B.2 Human Annotation Agreement

We assess the agreement between LLM judges and human preferences on 500 Qwen3-8B versus GPT-5.6 caption pairs, comprising 250 PixMo-Cap and 250 DenseFusion examples. Two annotators independently labeled each pair while blinded to the caption sources and to each other's judgments. Their annotations were reconciled into a final A/B/tie preference label for each pair. We measure exact agreement as the proportion of valid LLM judge responses that match these final labels; invalid responses are excluded from the denominator.

Table 7: Agreement between LLM judges and final human preferences on Qwen3-8B versus GPT-5.6 caption pairs.

LLM judge	Exact agreement
Gemini 3.1 Pro	429/499 (85.97%)
GPT-5.6 Sol	412/497 (82.90%)

Both judges show high agreement with the final human labels on this comparison set. These results support the use of our LLM-judge metric as a proxy for human caption preferences in this setting.

B.3 Training Reward Dynamics

Figure 5 compares the logged mean training reward and `train/frac_reward_zero_std` for RubiCap and MoCo Rubric. The latter is the fraction of rollout groups whose sampled captions all receive the same reward, yielding no within-group reward advantage for GRPO. Both runs contain 126 logged points over 624 update steps; step 600 is the checkpoint used for external evaluation.

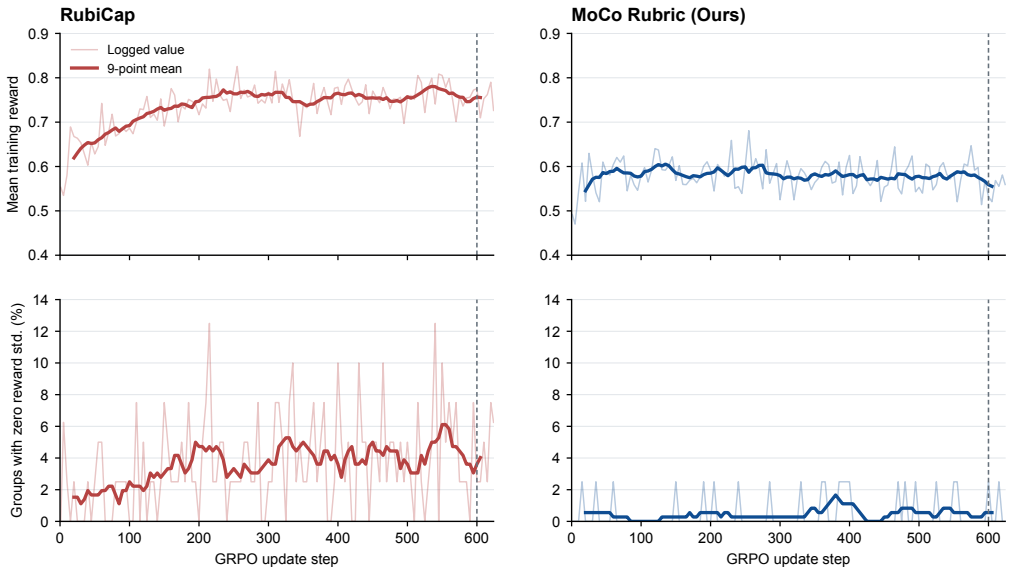


Figure 5: GRPO training dynamics of RubiCap and MoCo Rubric. The upper panels show mean training reward; the lower panels show the fraction of rollout groups with zero within-group reward standard deviation. Light traces are logged values and dark traces are centered nine-point moving averages. Dashed lines mark the evaluated step-600 checkpoints.

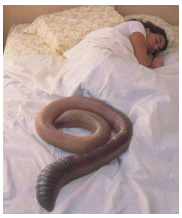
RubiCap's mean reward rises from 0.652 over steps 1–100 to 0.763 over steps 201–300, then remains near 0.758 over steps 301–600. MoCo Rubric does not show the same sustained rise toward a high-reward plateau: its mean reward is 0.588 over steps 201–300 and 0.577 over steps 301–600. Across logged points from steps 5–600, the mean fraction of zero-standard-deviation groups is 3.59% for RubiCap and 0.46% for MoCo Rubric; over steps 301–600, it is 4.33% and 0.58%, respectively. Thus, fewer rollout groups lack a within-group reward advantage under our method. These are single-run training diagnostics. The two methods use different rubrics and judges, and these runs start from different SFT checkpoints; reward magnitudes are therefore not directly comparable measures of caption quality, and the curves alone do not isolate the cause of the difference.

C Qualitative Case Studies



Figure 6: MoCo Rubric vs. Qwen3-VL-8B-Instruct. The selected examples illustrate differences in object pose, leaf shape, product layout, text alignment, and label position.

MoCo Rubric vs. RubiCap



MoCo Rubric

A person is lying on their side in a bed, covered by a white sheet, with their head resting on a pillow and one arm bent near their face. The bed has white bedding with visible wrinkles and folds, and a pale floral-patterned pillow or blanket is visible behind the person. In the foreground, a large segmented worm-like creature is curled on the bed, forming a loose ring with one end extending downward toward the lower left. The creature has a brownish, reddish-brown body with dark ridges and a ribbed, segmented texture, resembling a giant millipede or earthworm. Its body is thick and rounded, with a darker, more tapered tail-like end at the bottom. The scene has a humorous, surreal quality because the large creature appears to be resting beside the sleeping person, creating the impression that it is sharing the bed.

RubiCap

A person is lying on their side in a bed, covered with a white duvet or sheet, with their face obscured by a blur. They are wearing a white T-shirt and appear to be resting on a pillow with a light floral pattern. The bed has pale, wrinkled bedding and a beige wall or headboard behind it. At the foot of the bed, a large segmented worm-like creature is curled in a loose loop on the white sheets. It has a brownish, glossy, ribbed body with darker shading along the ridges, and its tapered tail curves downward toward the lower left. The creature's size is exaggerated compared to the bed, making it look like an enormous earthworm or centipede-like animal resting near the sleeper. The overall scene is humorous and surreal, as the sleeping person appears completely unaware of the large creature beside them.



MoCo Rubric

The image is a meme with a large bold black headline at the top on a white background that reads, "When you find out that Elon has left Twitter and switched to Reddit." Below the text is a movie-style image of a hooded, bearded figure standing dramatically in front of a bright glowing sunset or sunrise. The figure has long dark hair and beard, wears a dark hooded robe, and raises both hands outward in a commanding or theatrical gesture. Behind the figure is a large dark silhouette resembling a stylized Reddit alien mascot, with long curved arms or wing-like shapes extending upward and outward. The background is warm and hazy, with golden light on the right side and a misty landscape in the distance. Across the bottom of the image, large yellow text with a black outline reads, "Let the hunt for Elon Musk begin." The overall joke implies that when Elon Musk leaves Twitter and moves to Reddit, people will start searching for him on the platform.

RubiCap

The image is a meme with a white background and large bold black text at the top reading, "When you find out that Elon has left Twitter and switched to Reddit." Beneath the caption is a rectangular movie still showing a dark, hooded, bearded figure standing in a dramatic sunset or sunrise, with the face intentionally blurred. The figure has long dark hair and a long beard visible below the blur, wears a textured hooded garment, and raises both hands outward with fingers spread in a theatrical gesture. Behind the figure are large dark curved shapes resembling scythe blades or wing-like forms, silhouetted against a warm golden sky with a bright light source on the right. At the bottom of the still, large bold yellow text with a black outline reads, "Let the hunt for Elon Musk begin." The overall tone is humorous and conspiratorial, implying that Elon Musk's move to Reddit has triggered a dramatic, almost apocalyptic search for him.



MoCo Rubric

A square graphic with a solid dark burgundy background features centered white text in a clean sans-serif font. At the top, the heading reads "Hotel Service Tip:". Below it, the main message says "Be H.O.S.P.I.T.A.B.L.E. to guests" followed by the expanded acronym in parentheses: "(Help, Observe, Smile, Protect, Impress, Thank, Anticipate, Befriend, Listen, Engage)." Beneath the tip is a small tilde symbol followed by the attribution "The_Appraiser." In the bottom-right corner, the website "www.hotelevaluations.com.au" appears in smaller white text. The overall design is simple, minimal, and professional, resembling a social media tip card or hotel service reminder.

RubiCap

A square graphic with a solid dark burgundy background presents a hotel service tip in centered white text. At the top, in a larger bold sans-serif font, it reads "Hotel Service Tip:". Below it, in a larger bold block of text, the main message says: "Be H.O.S.P.I.T.A.B.L.E. to guests (Help, Observe, Smile, Protect, Impress, Thank, Anticipate, Befriend, Listen, Engage)." The acronym and its expanded meaning are arranged across multiple lines for readability. Near the lower center, a small tilde-like symbol "~" appears followed by the attribution "The_Appraiser" in white. In the bottom right corner, the website "www.hotelevaluations.com.au" is displayed in smaller white text. The overall design is minimal, with no images or decorative elements, relying entirely on typography and contrast between the white lettering and the deep red background.



MoCo Rubric

A square graphic with a muted lavender-purple background features large white text aligned to the left. At the top, in smaller bold white lettering, it reads "www.channelnewsasia.com." Below it, the main headline appears in very large bold white sans-serif type, broken across several lines: "Commentary: The fall of once-great Hyflux, a unicorn in the Singapore story." Under the headline, separated by a hyphen, the source is written in large bold white text as "- Channel NewsAsia," with "NewsAsia" on the next line. The design is minimal, with no images or decorative elements, relying entirely on typography and a solid pastel purple background.

RubiCap

The image is a square graphic with a muted lavender-purple background and large bold white text aligned to the left. At the top, in smaller white text, it reads "www.channelnewsasia.com." Below it, the main headline is displayed in very large, rounded, bold white lettering: "Commentary: The fall of once-great Hyflux, a unicorn in the Singapore story - Channel NewsAsia." The headline is broken across multiple lines, with "Commentary: The" on the first main line, "fall of once-great" on the next, "Hyflux, a unicorn in" on the following line, "the Singapore story" beneath that, and "- Channel NewsAsia" at the bottom. The typography is clean and modern, with wide spacing and a strong contrast between the bright white text and the soft purple background. There are no images, logos, or decorative elements besides the text.



MoCo Rubric

A dramatic promotional banner with a deep purple background features elegant white script text on the left reading "The Secret Love Life of Ophelia," with "The" and "of" in smaller italic lettering and "Secret Love Life" and "Ophelia" in large flowing cursive. On the right side, a person is shown from the shoulders up, wearing a rich purple blazer with structured shoulders and a matching purple garment underneath. Their hair is styled in tight braids or twists gathered upward, with warm golden-brown highlights visible against the dark purple backdrop. They are holding a bouquet of purple flowers, likely lisianthus or similar blossoms, with several open purple petals, pale green buds, and long green stems. The flowers partially frame the person's face, with some blooms near the nose and mouth area, creating a poetic, intimate composition. The person's expression is serious and intense, with dark eye makeup and a subtle, contemplative gaze. Long dangling earrings are visible on both sides, adding a refined detail. The overall mood is theatrical, romantic, and mysterious, with the purple color palette and floral imagery evoking themes of Shakespearean tragedy, hidden emotions, and the character of Ophelia.

RubiCap

A wide, moody promotional image with a deep purple background features a person on the right whose face is intentionally obscured by a large rectangular blur. The visible details suggest a formal portrait style: the person has dark, tightly braided hair with warm golden highlights, styled upward and back, and wears dangling silver earrings. They are dressed in a rich purple blazer with structured shoulders, creating a strong color harmony with the background. In front of them is a bouquet of purple lisianthus flowers, some fully bloomed and others still in bud form, with green stems and leaves. Several flowers are positioned close to the obscured face area, including a prominent purple bloom near the center and additional buds and stems extending to the left and right. The bouquet has a dramatic, theatrical feel, with the purple flowers and greenery contrasting against the dark purple suit and background. On the left side of the image, large elegant white script text reads "The Secret Love Life of Ophelia." The typography is ornate and romantic, with "The" and "of" in smaller italic lettering above and beside the larger words. "Secret Love Life" appears across the upper left and center-left, while "Ophelia" dominates the lower left in a very large, flowing serif style. The overall composition feels like a dramatic poster or book-cover design, using purple tones, floral imagery, and refined typography to evoke mystery, romance, and Shakespearean themes.

Figure 7: MoCo Rubric vs. RubiCap. The selected examples illustrate fabricated facial blurring and errors in text layout and typography.

MoCo Rubric vs. EvoLM



MoCo Rubric

The image is a meme-style graphic featuring a person in a dark suit, white dress shirt, and patterned tie standing in front of a solid blue curtain background. The person has a long gray-white beard and mustache, wears glasses, and has one arm raised high with a gloved hand, as if making a dramatic or emphatic point. The pose resembles a motivational speaker or political figure delivering a strong statement. Large bold white text with a black outline appears at the top and bottom of the image. The top text reads, "THE AMOUNT OF PEOPLE WHO DON'T KNOW THAT 'M&D' AND 'A&S' ARE EQUAL PRIORITY PAIRS IN P.E.M.D.A.S." The bottom text reads, "IS TOO DAMN HIGH." The overall tone is humorous and exaggerated, using the "too damn high" meme format to express frustration about people misunderstanding the order of operations in mathematics.

EvoLM

The image is a meme-style graphic featuring a person in a dark suit, white shirt, and patterned tie standing in front of a blue curtain background. The person's face is intentionally blurred, but their posture shows one arm raised high with a clenched fist, suggesting a dramatic or emphatic gesture. Large bold white text with a black outline appears across the top and bottom of the image. The top text reads, "THE AMOUNT OF PEOPLE WHO DON'T KNOW THAT 'M&D' AND 'A&S' ARE EQUAL PRIORITY PAIRS IN P.E.M.D.A.S." The bottom text reads, "IS TOO DAMN HIGH." The overall tone is humorous and exaggerated, using the "too damn high" meme format to express frustration about people misunderstanding the order of operations in mathematics.



MoCo Rubric

Three rectangular paper cards are pinned vertically on a brown corkboard background. Each card has a small clear plastic pushpin near the top center, holding it in place. The top card is cream-colored and written in large, thick black marker. It reads "OCT" on the first line and "28" below it. The middle card is light blue and also written in large black marker. It reads "NATE" on the first line and "BARGATZE" below it. The bottom card is bright pink and written in the same bold black marker style. It reads "FOO" on the first line and "FIGHTERS" below it. The overall image looks like a simple handmade reminder or event note pinned on a bulletin board, with the date, performer name, and band name arranged one card per line.

EvoLM

Three rectangular paper cards are pinned vertically on a brown corkboard background. Each card has bold, thick black handwritten lettering in a playful, uneven style. The top card is cream-colored and reads "OCT 28," with the date centered near the top. The middle card is light blue and reads "NATE BARGATZE," with the name written across the card in large, expressive letters. The bottom card is bright pink and reads "FOO FIGHTERS," with the band name written in large black marker. Each card is attached to the corkboard with a small clear plastic pushpin near the top center, creating a simple event-planning or reminder-style display.



MoCo Rubric

A square promotional poster shows a close-up, cropped view of a shirtless person from the lower face down to the upper torso, with the head and neck area partially cut off at the top edge. The visible skin is softly lit with warm tones, and the shoulders, collarbones, chest, and upper arms are centered in the frame. The person's arms appear relaxed, with one arm bent across the lower right side and the other resting near the left side, creating a gentle, posed composition. The background is dark and blurred, giving the image a moody, cinematic feel. Large decorative serif text dominates the center of the image, reading "DRAG BECOMES HIM" in dark brown, ornate capital letters with curled flourishes and a subtle shadow effect. Beneath it, smaller uppercase text reads "A DOCUMENTARY ABOUT JINKX MONSOON," followed by "COMING SUMMER 2014." The typography has a vintage, theatrical style that matches the drag documentary theme. The overall design is elegant and dramatic, with the bold title contrasting against the softly lit human form and dark background.

EvoLM

A square promotional poster shows a close-up, cropped view of a shirtless person from the shoulders up, with the face intentionally blurred and the lower body cut off at the waist. The person's bare chest and arms are visible, with soft lighting and a dark, moody background. Centered over the torso is large decorative serif text reading "DRAG BECOMES HIM" in dark brown, shadowed lettering. Beneath it, smaller text reads "A DOCUMENTARY ABOUT JINKX MONSOON" and "COMING SUMMER 2014." The overall design has a dramatic, cinematic feel, using warm skin tones against a dark backdrop and ornate typography to create a theatrical documentary poster.



MoCo Rubric

A rectangular cutting board is shown against a plain white background, viewed from above. The board has a light tan, cork-like surface with a speckled, grainy texture across the entire face. It has softly rounded corners and a slightly curved top and bottom edge, giving it a gentle, ergonomic shape. Near the upper center of the board, the word "epicurean" is printed in small dark text. Four dark red corner reinforcements are visible, one at each corner, with rounded edges that match the board's shape. The red corner pieces are slightly raised or bordered, creating a subtle contrast with the tan surface. On the upper left corner, there is a large circular cutout or hole, positioned near the top-left area, likely for hanging or carrying. The overall design is simple and functional, with a natural-looking cutting board surface and red corner accents.

EvoLM

A rectangular cutting board is shown against a plain white background, viewed from above. The board has a light tan, speckled surface with a slightly textured, cork-like or composite material appearance. It has rounded corners and a subtle curved top edge, giving it a soft, ergonomic shape. Near the upper center of the board, the brand name "epicurean" is printed in small dark text. Along the top and bottom edges are darker reddish-brown border strips, and matching reddish-brown corner reinforcements appear at the top-left, top-right, bottom-left, and bottom-right corners. On the upper left side, there is a large circular cutout or hole, likely for hanging or carrying. The overall design is simple and functional, with a neutral beige surface and dark red accent corners.



MoCo Rubric

A wide banner-style image shows a close-up of a white lotus flower in bloom against a soft, blurred green background of leaves and foliage. The flower is centered slightly below the upper half of the image, with its pale cream petals gently unfurling outward in layered curves. The petals have fine vertical ridges and subtle shadows that give them a delicate, textured appearance, while the central bud remains rounded and softly illuminated. Large white uppercase text overlays the image. Across the top, the text reads "A GUIDE TO PLANNING FOR," though the word "PLANNING" appears misspelled as "PLANNIG." Near the bottom, the text reads "AGED CARE" in very large bold white letters. The typography is clean, modern, and sans-serif, with generous spacing and a strong contrast against the green background. A small green stem rises from the flower toward the lower center. In the lower left corner, part of another white flower is visible but out of focus, adding depth to the scene. The overall mood is calm, natural, and serene, with the lotus symbolizing purity, growth, and care.

EvoLM

A wide banner-style image shows a close-up of a white lotus flower in bloom against a soft, blurred green background of leaves and water. The flower is centered slightly to the right, with delicate white petals layered around a rounded, pale yellow-green center, giving it a fresh, serene appearance. Large bold white uppercase text overlays the image, reading "A GUIDE TO PLANNING FOR" across the top and "AGED CARE" across the bottom. The text is clean and modern, with generous spacing, and the overall design suggests a calm, thoughtful guide about planning for aged care.

Figure 8: MoCo Rubric vs. EvoLM. The selected examples illustrate fabricated facial blurring and border details, as well as errors in text layout and transcription.

D Proofs and Additional Momentum Analysis

D.1 Proof of Proposition 1

At synchronization boundary s, the momentum update is

$$\bar{\phi}_s = m\bar{\phi}_{s-1} + (1-m)\bar{\theta}_s. \tag{5}$$

Define

$$d_s = \bar{\theta}_s - \bar{\phi}_s, \qquad \xi_s = \bar{\theta}_s - \bar{\theta}_{s-1}, \tag{6}$$

so that $e_s = |d_s|$ and $\Delta_s = |\xi_s|$. Substituting Eq. (5) into d_s gives

$$\begin{aligned} d_s &= \bar{\theta}_s - \left[m\bar{\phi}_{s-1} + (1-m)\bar{\theta}_s\right] \\ &= m(\bar{\theta}_s - \bar{\phi}_{s-1}) \\ &= m(\xi_s + d_{s-1}). \end{aligned} \tag{7}$$

Taking norms and applying the triangle inequality yields

$$e_s \le m e_{s-1} + m\Delta_s. \tag{8}$$

Recursively expanding this inequality gives

$$e_s \le m^s e_0 + \sum_{j=1}^{s} m^{s-j+1}\Delta_j. \tag{9}$$

Because synchronization occurs every H Policy updates and each update satisfies $|\theta_{t+1} - \theta_t| \le \delta$, we have

$$\begin{aligned} \Delta_s &= |\bar{\theta}_s - \bar{\theta}_{s-1}| \\ &\le \sum_{h=1}^{H} \left|\theta_{(s-1)H+h} - \theta_{(s-1)H+h-1}\right| \\ &\le H\delta. \end{aligned} \tag{10}$$

Substituting this result into Eq. (9) and evaluating the geometric series yields

$$\begin{aligned} e_s &\le m^s e_0 + H\delta \sum_{j=1}^{s} m^{s-j+1} \\ &= m^s e_0 + \frac{m(1-m^s)}{1-m} H\delta. \end{aligned} \tag{11}$$

Since shared initialization gives $\bar{\theta}_0 = \bar{\phi}_0$ and therefore $e_0 = 0$,

$$e_s \le \frac{m(1-m^s)}{1-m} H\delta \le \frac{mH\delta}{1-m}. \tag{12}$$

For the momentum-update magnitude, Eq. (5) gives

$$\begin{aligned} \bar{\phi}_s - \bar{\phi}_{s-1} &= (1-m)(\bar{\theta}_s - \bar{\phi}_{s-1}) \\ &= (1-m)(\xi_s + d_{s-1}). \end{aligned} \tag{13}$$

Taking norms gives

$$u_s = |\bar{\phi}_s - \bar{\phi}_{s-1}| \le (1-m)(e_{s-1} + \Delta_s), \tag{14}$$

which completes the proof. □

D.2 Local Drift–Noise Analysis

We next derive the stochastic smoothing result used in the main text. Consider the local Policy-displacement model

$$\xi_s = \bar{\theta}_s - \bar{\theta}_{s-1} = Hv + \varepsilon_s, \tag{15}$$

where

$$\mathbb{E}[\varepsilon_s] = 0, \qquad \mathrm{Cov}(\varepsilon_s) = H\Sigma. \tag{16}$$

We assume that v is approximately constant over the local analysis window and that ε_s is independent across non-overlapping synchronization intervals. Because d_{s-1} depends only on noise from preceding intervals, this assumption also implies that ε_s is independent of d_{s-1}.

Substituting Eq. (15) into Eq. (7) gives

$$d_s = md_{s-1} + mHv + m\varepsilon_s. \tag{17}$$

Under shared initialization, $d_0 = 0$. Recursively expanding Eq. (17) yields

$$d_s = mH \sum_{j=1}^{s} m^{s-j} v + m \sum_{j=1}^{s} m^{s-j} \varepsilon_j. \tag{18}$$

Its finite-step mean is

$$\mathbb{E}[d_s] = \frac{mH(1-m^s)}{1-m} v. \tag{19}$$

Using the independence of the noise terms, its covariance is

$$\begin{aligned} \mathrm{Cov}(d_s) &= m^2 \sum_{j=1}^{s} m^{2(s-j)} \, \mathrm{Cov}(\varepsilon_j) \\ &= \frac{m^2 H(1-m^{2s})}{1-m^2} \Sigma. \end{aligned} \tag{20}$$

For $0 \le m < 1$, taking $s \to \infty$ gives

$$\mathbb{E}[d_s] \longrightarrow \frac{mH}{1-m} v, \qquad \mathrm{Cov}(d_s) \longrightarrow \frac{m^2 H}{1-m^2} \Sigma. \tag{21}$$

Thus, the stationary expected squared tracking error is

$$\mathbb{E}\left[|d_s|^2\right] = \frac{m^2 H^2}{(1-m)^2} |v|^2 + \frac{m^2 H}{1-m^2} \, \mathrm{tr}(\Sigma). \tag{22}$$

We now consider the momentum-model update

$$q_s = \bar{\phi}_s - \bar{\phi}_{s-1}. \tag{23}$$

From Eq. (13),

$$q_s = (1-m) \left(d_{s-1} + Hv + \varepsilon_s \right). \tag{24}$$

Its finite-step mean is

$$\mathbb{E}[q_s] = H(1-m^s)v. \tag{25}$$

Since d_{s-1} and ε_s are independent,

$$\begin{aligned} \mathrm{Cov}(q_s) &= (1-m)^2 \left[\mathrm{Cov}(d_{s-1}) + H\Sigma \right] \\ &= \frac{1-m}{1+m} (1-m^{2s}) H\Sigma. \end{aligned} \tag{26}$$

Taking the stationary limit gives

$$\mathbb{E}[q_s] \longrightarrow Hv, \qquad \mathrm{Cov}(q_s) \longrightarrow \frac{1-m}{1+m} H\Sigma. \tag{27}$$

For periodic full copying, $m = 0$, so $q_s = \xi_s$ and $\mathrm{Cov}(q_s) = H\Sigma$. EMA therefore reduces the stationary update covariance, relative to full copying at the same H, by the factor

$$\frac{1-m}{1+m}. \tag{28}$$

However, Eq. (21) shows that increasing m also increases the mean tracking lag and tracking variance. Similarly, increasing H increases both the systematic displacement and accumulated noise between synchronization boundaries. These results establish a parameter-space tracking–stability trade-off; they do not imply that Policy optimization necessarily improves the task-level quality of rubric generation or judging.