CuRe: Claim-Level Rubric Rewards for Video Captioning

Anonymous authors
Paper under double-blind review

Abstract

Existing reward functions for dense video captioning largely assume reference captions provide complete supervision, overlooking that a single reference can rarely cover all valid visual content in a video. To address this issue, we propose Claim-Level Rubric Rewards (CuRe), a reward framework that decomposes captions into visual claims such as entities and actions, and verifies each claim directly against the video. This claim-level verification enables CuRe to distinguish fine-grained errors from valid details omitted by the reference. We then define a reference-anchored credit weighting scheme that assigns credit based on both visual support and reference alignment: reference-matched claims receive full credit, video-supported claims beyond the reference receive bounded credit, and unsupported claims receive zero credit. This design rewards visually grounded coverage rather than strict reference matching in text, allowing valid details beyond the reference to contribute while keeping unsupported claims penalized. On three public captioning benchmarks, CuRe achieves the strongest results among open-source models. To test whether CuRe-generated captions provide better supervision than the original captions, we re-caption three video-caption corpora (867K samples) and pretrain VLMs with either the original or CuRe-generated captions. CuRe-recaptioned data improves performance on 25 of 27 metrics across six downstream video-understanding benchmarks, demonstrating that CuRe-generated captions provide stronger video-language supervision.

1 Introduction

Video captioning is a core task in video understanding (Maaz et al., 2024; Yang et al., 2024) and supports downstream applications from controllable video generation (Xiong et al., 2024) to content recommendation. Building detailed and reliable captioning models typically relies on Supervised Fine-Tuning (SFT) with high-quality reference captions distilled from proprietary models or human annotation (Yuan et al., 2025; Chen et al., 2024). However, each reference provides only one possible account of a video, and SFT learns to imitate these text examples rather than to ground descriptions in visual facts (Arora et al., 2022; Song & Zheng, 2026). This limitation is more pronounced in dense captioning, where a single reference rarely covers all valid visual content.

Reinforcement learning (RL) offers a promising alternative by optimizing the captioner against reward signals derived from its own rollouts (Meng et al., 2025; Zhong et al., 2025; Xing et al., 2026). However, designing an effective reward function for video captioning remains highly challenging. Existing approaches commonly derive rewards from holistic scoring of caption quality (Ye et al., 2025; Xing et al., 2026) or from aligning generated captions with reference descriptions via lexical metrics(e.g., ROUGE, CIDEr) (Pasunuru & Bansal, 2017; Oliveira dos Santos et al., 2021) and model judgments (Meng et al., 2025; Zhong et al., 2025). While both paradigms report consistent gains over SFT, neither provides the fine-grained, visually grounded feedback that video captioning demands. Specifically, they leave two fundamental challenges largely unaddressed: i) Holistic scoring collapses diverse visual dimensions into a single scalar, making it difficult to distinguish fine-grained errors from valid descriptions; ii) A single reference can rarely capture all valid visual content, so strictly matching rewards against it cannot tell whether an unmatched statement is false or describes valid visual content that is present in the video but omitted from the reference.

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.

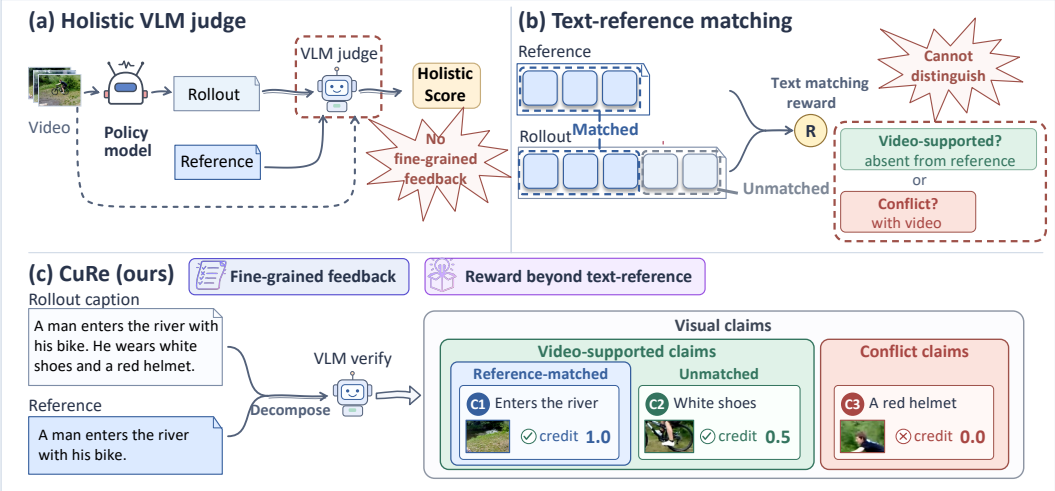


Figure 1: (a) **Holistic VLM judges** provide only coarse whole-caption feedback. (b) **Text-reference matching** cannot distinguish valid details absent from the reference from conflicts with the video. (c) **CuRe** decomposes captions into visual claims and verifies them directly against the video. Reference-anchored credit weighting scheme then rewards reference-matched claims fully, valid residual claims partially, and unsupported claims with zero credit.

To address these issues, we propose **C**laim-Level R**u**bric **Re**wards (**CuRe**), a reward framework that supervises caption training by jointly rewarding factual accuracy and semantic coverage.

First, we establish a structured rubric to decompose generated captions into visual claims across ten categories (e.g., entities, actions, and spatial relations). By grounding each claim in the video, CuRe replaces intractable holistic judgments with simpler, precise checks, directly addressing challenge (i) by catching specific errors overlooked by holistic models. Second, we extend the reward beyond reference matching. Rather than treating the reference as a closed coverage target, it uses reference claims to identify salient content and visual evidence to determine whether additional details are valid. We implement this through a *reference-anchored credit weighting* scheme: claims matched to the reference receive full credit, video-supported claims beyond the reference receive bounded credit, and unsupported claims receive zero credit. The reference thus serves as an anchor for prioritizing salient content, while the video provides the evidence needed to reward valid details that the reference omits. This directly addresses challenge (ii) by allowing such details to contribute to the reward, while the zero-credit rule prevents reward hacking through verbose elaboration.

Extensive experiments across three distinct settings show that the CuRe-trained captioner outperforms open-source models. On direct captioning benchmarks, CuRe achieves the highest overall score on EventHallusion, the best results across all three VCapsBench metrics, and the highest precision and F1 on DREAM-1K among the evaluated open-source models. Within the Prism framework for caption quality evaluation (Qiao et al., 2024), CuRe outperforms substantially larger baselines like Qwen3VL-A22B-235B. Furthermore, to test whether CuRe-generated captions provide better supervision than the original ones, we re-caption three video-caption corpora (totally 867K samples) and pretrain VLMs with either the original or CuRe-generated captions. CuRe-recaptioned data achieves gains on 25 of 27 metrics across six downstream video-understanding benchmarks, demonstrating that CuRe captions provide stronger video-language supervision.

Our key contributions can be summarized as follows:

- We propose **CuRe**, a claim-level rubric reward framework that decomposes video captions into visual claims. By grounding each claim in the video, CuRe provides precise, per-claim feedback and effectively penalizes fine-grained factual errors.
- We introduce a *reference-anchored credit weighting* scheme that extends the reward beyond reference matching. The reference caption provides the main reference, while the video supplies additional details, together yielding broader coverage than reference matching alone.
- Extensive experiments across three settings demonstrate that CuRe outperforms state-of-the-art methods and boosts downstream training performance.

2 Related Work

2.1 Video Captioning

Video captioning has shifted from producing concise summaries to generating detailed, multi-axis descriptions. Early paradigms predominantly rely on supervised fine-tuning, leveraging either human-annotated datasets (Xu et al., 2016; Chen & Dolan, 2011; Wang et al., 2019) or distilled corpora from proprietary models (Chen et al., 2024; Zhang et al., 2024). To further scale up supervision, recent methods employ fine-tuned captioners to self-generate massive training data to explore the model capacity limits (Wang et al., 2024; Yuan et al., 2025; Cho et al., 2025). More recently, reinforcement learning (RL) has emerged as a promising alternative. Several pioneering works (Ahn et al., 2024; Tang et al., 2025; Meng et al., 2025) have integrated advanced RL paradigms (e.g., PPO, DPO, and GRPO) into video captioning task. However, these methods typically rely on coarse-grained feedback computed from the entire caption, which obscures fine-grained errors within it.

2.2 Rubric Rewards for Reinforcement Learning

Rubric-based rewards are effective for reinforcement learning on open-ended generation tasks, where holistic scalar rewards often fail to capture diverse quality dimensions. Instead of relying on single-score judgments, rubric rewards decompose evaluation into explicit and independently assessable criteria, providing more structured supervision signals for policy optimization. Recent works explore rubric generation and scalable reward construction for open-ended RL, including Chain-of-Rubrics in RM-R1 (Chen et al., 2026), dynamically generated checklist rewards in RaR (Gunjal et al., 2026), and scalable rubric synthesis frameworks such as OpenRubrics, Auto-Rubric, and RL-CER (Liu et al., 2026; Xie et al., 2026; Sheng et al., 2026). More recently, rubric-guided RL has been extended to multimodal generation, where RubiCap applies rubric-based rewards to dense image captioning (Huang et al., 2026). However, when visual claims fall outside the rubric's coverage, rubric satisfaction alone cannot distinguish unsupported statements from valid additional details.

3 Methodology

In this section, we describe CuRe, a three-stage claim-level reward pipeline, as illustrated in Figure 2. Given a rollout caption and a reference caption, the pipeline first decomposes both into visual claims, verifies candidate claims against the video, and matches them to reference claims (§3.1). Based on these claim-level verification and matching results, the second stage assigns credit to each claim through a reference-anchored weighting scheme that balances visual grounding with reference alignment (§3.2). The resulting category-wise credit scores are then aggregated into a single scalar and combined with a length-control penalty to form the final reward (§3.3).

3.1 Claim-Level Rubric Construction

Given a video x and a model-rollout caption y, we use an LLM-based decomposer (Qwen Team, 2025) to project the caption into a set of typed candidate claims:

$$\mathcal{C} = \mathrm{Decompose}(y) = \{(c_i, k_i)\}_{i=1}^{n}, \tag{1}$$

where c_i is a visual claim, $k_i \in \mathcal{K}$ denotes its visual category under the schema illustrated in Appendix B, and n is the number of extracted claims. For each category $k \in \mathcal{K}$, we write $\mathcal{C}_k = \{c_i \mid (c_i, k_i) \in \mathcal{C}, k_i = k\}$ to denote the candidate claims assigned to k. We also decompose a reference caption $y^\star$ generated by an advanced proprietary VLM (Pichai et al., 2025), yielding the reference claim set $\mathcal{C}^\star = \mathrm{Decompose}(y^\star)$.

CuRe then verifies and matches claims within each visual category. For each category k, let $\mathcal{C}_k^\star$ denote the reference claims assigned to k. A VLM (Qwen Team, 2025) video verifier checks each candidate claim in $\mathcal{C}_k$ against the video and retains the supported subset:

$$\mathcal{S}_k = \{c \in \mathcal{C}_k \mid v(x, c) = 1\}, \tag{2}$$

where $v(x, c)$ is the VLM verifier indicating whether claim c is supported by video x. That is, $\mathcal{S}_k \subseteq \mathcal{C}_k$ contains the candidate claims in category k that are supported by the video. The supported

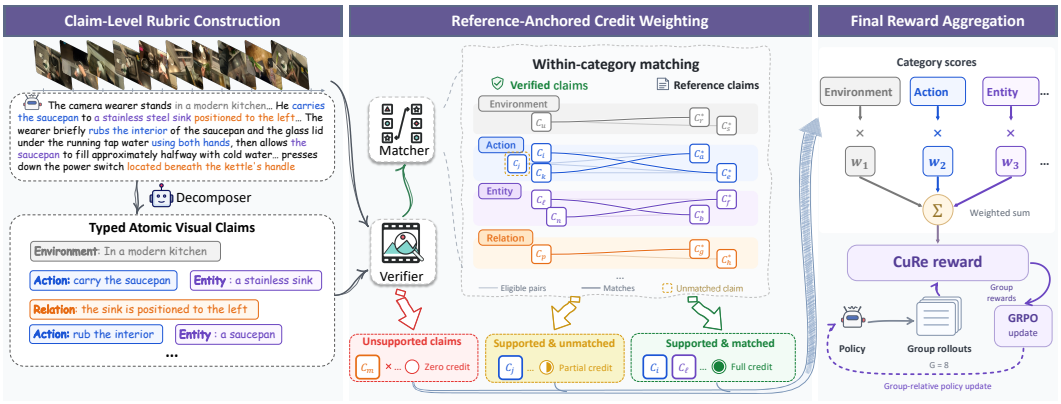


Figure 2: Overview of the CuRe framework. CuRe decomposes captions into atomic visual claims, verifies them against the video, and matches supported claims one-to-one with reference claims within each category. It then applies reference-anchored credit weighting to assign corresponding credit. The calibrated category-wise scores are finally aggregated into the overall reward.

candidate claims $\mathcal{S}_k$ are then matched to reference claims $\mathcal{C}_k^\star$ within the same category:

$$\mathcal{M}_k = \{(c, c^\star) \in \mathcal{S}_k \times \mathcal{C}_k^\star \mid m(c, c^\star) = 1\}, \tag{3}$$

where $c^\star \in \mathcal{C}_k^\star$ is a reference claim, and $m(c, c^\star)$ indicates that the two claims are semantically aligned. Each claim can appear in at most one matched pair.

We define base claim-level precision and recall scores:

$$p_k = \frac{|\mathcal{S}_k|}{|\mathcal{C}_k|}, \qquad r_k = \frac{|\mathcal{M}_k|}{|\mathcal{C}_k^\star|}. \tag{4}$$

Here, precision measures the visual grounding of the generated caption, while recall measures how much of the reference claim set is covered by the supported candidate claims.

3.2 Reference-Anchored Credit Weighting

The base claim-level scores in Eq. 4 provide a simple starting point for measuring visual grounding and reference coverage. However, reference claims are reliable but incomplete: they provide strong anchors for salient content, but should be treated as a partial view rather than a complete account of the video's visual evidence. This motivates a reference-anchored credit weighting scheme, which assigns preferential credit to reference-matched claims while allowing claims that are video-supported but unmatched to the reference to contribute in a controlled manner.

For precision, CuRe replaces uniformly support-based credit with reference-aware credit. Reference-matched supported claims ($\mathcal{M}_k$) receive full credit, while unmatched supported claims ($\mathcal{S}_k \setminus \mathcal{M}_k$) receive attenuated credit $w_u \in [0, 1]$:

$$p_k^{\text{CuRe}} = \frac{|\mathcal{M}_k| + w_u \left(|\mathcal{S}_k| - |\mathcal{M}_k|\right)}{|\mathcal{C}_k|}. \tag{5}$$

This favors reference-matched claims while retaining discounted credit for additional video-supported details. For recall, CuRe avoids treating the reference claims as a closed coverage target. Unmatched supported claims may expand the effective target, but only through discounted credit:

$$r_k^{\text{CuRe}} = \frac{|\mathcal{M}_k| + w_u \min\left(|\mathcal{S}_k| - |\mathcal{M}_k|,\ \delta_k\right)}{|\mathcal{C}_k^\star| + w_u \min\left(|\mathcal{S}_k| - |\mathcal{M}_k|,\ \delta_k\right)}, \tag{6}$$

where $\delta_k = |\mathcal{C}_k^\star| - |\mathcal{M}_k|$ caps the additional credit from unmatched supported claims at the number of uncovered reference claims. This cap prevents the captioning model from hacking the recall metric by generating video-supported but redundant visual details not covered by the reference caption.

3.3 Final Reward Aggregation

After obtaining the category-wise credit scores, CuRe aggregates precision and recall over their respective active categories. Specifically, precision is computed over non-empty candidate categories $\mathcal{K}_{\text{cand}} = \{k \in \mathcal{K} : |\mathcal{C}_k| > 0\}$ and recall over non-empty reference categories $\mathcal{K}_{\text{ref}} = \{k \in \mathcal{K} : |\mathcal{C}_k^\star| > 0\}$. Weighting each category by its prior ω_k, we define the aggregated metrics as:

$$p^{\text{CuRe}} = \frac{\sum_{k \in \mathcal{K}_{\text{cand}}} \omega_k p_k^{\text{CuRe}}}{\sum_{k \in \mathcal{K}_{\text{cand}}} \omega_k}, \qquad r^{\text{CuRe}} = \frac{\sum_{k \in \mathcal{K}_{\text{ref}}} \omega_k r_k^{\text{CuRe}}}{\sum_{k \in \mathcal{K}_{\text{ref}}} \omega_k}. \tag{7}$$

We set $p^{\text{CuRe}} = 0$ for captions with no extracted claims ($\mathcal{K}_{\text{cand}} = \varnothing$). We define the overall CuRe reward as a weighted combination of the aggregated precision and recall:

$$R^{\text{CuRe}} = \lambda\, p^{\text{CuRe}} + (1 - \lambda)\, r^{\text{CuRe}}, \tag{8}$$

where $\lambda \in [0, 1]$ balances precision against recall: larger λ emphasizes precision, favoring captions that avoid unsupported claims, while smaller λ emphasizes recall, favoring broader coverage of visual content. To discourage over-generation, we subtract a bounded length penalty Ω based on the ratio of generated to reference claim counts, with details provided in Appendix A.

For training, we adopt the standard GRPO (Shao et al., 2024) framework with CuRe as the reward function. For each video, the policy model samples a group of eight captions, which are scored by evaluating its candidate claims against the video, together with the length-control penalty. This claim-level reward provides fine-grained and coverage-aware feedback, steering optimization toward captions that are visually grounded and informative rather than verbose but unsupported.

4 Experiments

In this section, we evaluate CuRe from three perspectives: direct caption quality on open-ended captioning benchmarks (§4.2), downstream task performance after CuRe-enhanced pretraining (§4.3), and caption-to-QA evaluation across multiple benchmarks (§4.4). We further analyze the contribution of each reward component through ablations (§4.5). We also assess the accuracy of the reward models in claim decomposition, video verification, and semantic matching (Appendix C.5). We complement these quantitative evaluations with qualitative caption comparisons in Appendix E.

4.1 Experimental Setup

Implementation Details. We initialize the policy captioner from Qwen3-VL-30B-A3B-Instruct (Qwen Team, 2025) and post-train it on a Gemini-3.0-Pro-distilled caption corpus drawn from open-source web videos (Yuan et al., 2025). The captioner is first trained with a 78,144-sample video caption dataset as SFT warm-start, and then optimized on a deduplicated 8,000-video subset with GRPO. In the reward runtime, Qwen3-VL-235B-A22B-Instruct decomposes only the rollout captions, verifies rollout claims against the video, and performs the one-to-one matcher under task-specific prompts (Appendix F). Gemini-3.0-Pro (Pichai et al., 2025) provides the reference captions and their decomposed reference claims $\mathcal{C}^\star$. We then perform one semantic match between supported rollout claims and these reference claims. We set $w_u = 0.5$ and use $\lambda = 0.7$ in the main experiments to prioritize precision and discourage unsupported claims. Detailed training settings are provided in Appendix A. None of the SFT or GRPO videos overlap with our benchmarks.

Baselines. To show the leading performance of CuRe, we compare our captioner not only with open-source video captioners such as Tarsier (Wang et al., 2024), LLaVA-Video (Zhang et al., 2024), Tarsier2 (Yuan et al., 2025), and OwlCap (Zhong et al., 2025), but also with general-purpose VLMs including Qwen2.5-VL, InternVL3.5, and Qwen3-VL (Qwen Team, 2025). We also include proprietary models, including Seed 2.0 Pro and Gemini 3.0 Pro, as reference baselines to assess how far CuRe narrows the gap between open-source video captioners and leading proprietary VLMs.

4.2 Caption Quality Evaluation

In this part, we evaluate CuRe on three public benchmarks for video caption quality.

Table 1: Evaluation on captioning and hallucination benchmarks. All scores are percentages. EventHallusion settings cover rare-event recognition (Entire), temporal-switch hallucination (Inter.), and prompt-misleading robustness (Mislead.). VCapsBench reports accuracy (AR), inconsistency (IR, lower is better), and coverage (CR) rates. DREAM-1K reports precision (Prec.), recall (Rec.), and F1. **Bold** and underline mark the best and second-best open-source results.

Model	EventHallusion (Description)				VCapsBench			DREAM-1K		
	Entire↑	Inter.↑	Mislead.↑	Overall↑	AR↑	IR↓	CR↑	Prec.↑	Rec.↑	F1↑
Proprietary models										
Seed 2.0 Pro (ByteDance Seed, 2026)	64.22	76.17	96.84	77.83	60.37	14.59	70.68	32.97	50.80	39.99
Gemini 3.0 Pro (Pichai et al., 2025)	65.13	68.39	98.04	74.59	70.29	15.44	83.13	42.67	39.26	40.89
Open-source models										
LLaVA-Video-7B (Zhang et al., 2024)	27.50	32.60	75.50	41.40	45.11	17.46	54.65	37.90	28.40	32.50
OwlCap-7B (Zhong et al., 2025)	20.18	51.81	80.00	49.87	43.82	15.79	52.04	34.10	35.30	34.70
Tarsier2-7B (Yuan et al., 2025)	**54.60**	53.10	93.70	63.30	38.51	13.89	44.73	41.50	38.80	40.10
InternVL3.5-30B-A3B (Wang et al., 2025)	34.86	54.40	75.79	54.16	56.85	17.43	68.86	39.07	28.17	32.74
Qwen3-VL-30B-A3B-Instruct (Qwen Team, 2025)	32.11	60.10	87.37	58.94	61.26	14.30	71.49	33.35	38.32	35.66
Tarsier-34B (Wang et al., 2024)	38.50	40.40	83.20	50.10	-	-	-	41.40	32.40	36.30
Qwen2.5-VL-72B-Instruct (Bai et al., 2025)	24.77	62.69	81.05	56.68	61.20	16.02	72.87	34.72	28.90	31.54
Qwen3-VL-235B-A22B-Instruct (Qwen Team, 2025)	39.45	65.28	81.05	61.96	64.80	16.09	77.23	29.48	**40.77**	34.22
CuRe	45.87	**69.95**	**95.79**	**69.52**	**67.77**	**13.78**	**78.60**	**48.21**	36.26	**41.39**

Settings. EventHallusion (Zhang et al., 2025) assesses event hallucination under the Entire, Interleave, and Misleading settings, using GPT-based description matching to check whether captions are consistent with the target event. We report description-level scores. VCapsBench (Zhang et al., 2026) evaluates how well captions support text-only QA using human-verified video questions with ternary answers. It reports Accuracy Rate (AR), Inconsistency Rate (IR, lower is better), and Coverage Rate (CR). DREAM-1K (Wang et al., 2024) uses AutoDQ to extract events from generated and reference captions and compute event-level precision, recall, and F1.

Results and analysis. As shown in Table 1, CuRe achieves the strongest overall direct-captioning performance among the evaluated open-source models. On EventHallusion, CuRe obtains an Overall score of 69.5, exceeding its same-sized Qwen3-VL-30B-A3B-Instruct baseline by 10.6 points and the strongest open-source baseline, Tarsier2-7B, by 6.2 points. On VCapsBench, CuRe ranks first among open-source models on all three metrics, with broader coverage and fewer inconsistencies. It also outperforms the much larger Qwen3-VL-235B-A22B-Instruct and the proprietary Seed 2.0 Pro across all three metrics. On DREAM-1K, CuRe achieves the highest precision and F1 among open-source models, although its recall is 4.5 points below Qwen3-VL-235B-A22B-Instruct. Overall, CuRe demonstrates remarkable performance gains in factual accuracy (EventHallusion Overall, VCapsBench IR, and DREAM-1K precision), while maintaining competitive semantic coverage (VCapsBench CR and DREAM-1K recall). This suggests that CuRe favors faithful and informative event descriptions over recall gains from potentially noisy or redundant details.

4.3 Downstream Evaluation via CuRe Pretraining

In this part, we use model-generated video captions for end-to-end VLM pretraining, with downstream benchmark performance serving as a proxy for caption quality. This setup tests whether CuRe provides stronger supervision for video-language learning across different caption corpora.

Settings. We initialize the model with a Qwen3-4B (Yang et al., 2025) language backbone, the Qwen3 vision encoder (Qwen Team, 2025), and a randomly initialized projector, and train it with a three-stage protocol following ShareGPT4V (Chen et al., 2025; 2024) and CapRL (Xing et al., 2026). Stage 1 performs image-text alignment on BLIP-3-KALE (Awadalla et al., 2024). Stage 2 continues pretraining on Molmo2-Cap, Tarsier2-Recap (Yuan et al., 2025), or LLaVA-Video-Caption. We reannotate each corpus with CuRe and compare models pretrained on the original and reannotated captions. Stage 3 applies SFT on a fixed VideoQA mixture from Molmo2-AMA (Clark et al., 2026) and LLaVA-Video (Zhang et al., 2024). Training details are provided in Appendix D.

Benchmarks. We evaluate downstream video-language performance on six complementary benchmarks: MMVU for expert-level video understanding and reasoning (Zhao et al., 2025), MVBench for temporal perception and cognition (Li et al., 2024), MotionBench for fine-grained motion understanding (Hong et al., 2025), TOMATO for visual temporal reasoning (Shangguan et al., 2025), and

Table 2: Evaluation of downstream VLM performance using original or CuRe-generated video captions for pretraining. "CuRe Recap." replaces the original captions for the same source videos. All other training and evaluation settings are held fixed within each pair. **Bold** marks the better result.

Video Dataset	Caption Source	MMVU	MVBench	MotionBench	PerceptionTest	TOMATO	VideoMME			
							Overall	Long	Medium	Short
Molmo2-Cap (104k)	Original	**38.10**	45.10	47.98	48.99	22.64	52.15	42.33	49.33	64.78
	CuRe Recap.	37.40	**48.83**	**49.38**	**52.30**	**26.08**	**54.07**	**44.22**	**50.89**	**67.11**
Tarsier2-Recap (585k)	Original	25.40	48.85	49.00	51.84	25.00	51.52	40.78	48.78	65.00
	CuRe Recap.	**28.70**	**49.48**	**50.42**	**53.15**	**26.68**	**53.70**	**42.56**	**51.67**	**66.89**
LLaVA-Video-Caption (178k)	Original	31.20	**49.08**	48.86	52.68	22.57	52.30	39.00	51.56	66.33
	CuRe Recap.	**37.30**	48.73	**50.32**	**53.25**	**24.93**	**54.37**	**42.22**	**52.00**	**68.89**

Table 3: Caption-to-QA evaluation of different video captioners under the Prism protocol. Seed 2.0 Pro answers benchmark questions using only the video description generated by each captioner. **Bold** and underline mark the best and second-best open-source results, respectively.

Captioner	VideoMME				MMVU			MVBench	MotionBench	TOMATO	PerceptionTest
	Overall	Short	Medium	Long	MCQ	Open	Overall	Overall	Overall	Score	Overall
Proprietary captioners											
Seed 2.0 Pro	66.63	74.67	68.00	57.22	78.24	55.73	69.80	56.23	53.19	35.58	65.75
Gemini 3.0 Pro	68.52	75.11	69.67	60.78	77.28	58.13	70.10	59.13	50.85	35.31	63.73
Open-source captioners											
Tarsier2-7B	33.19	40.44	30.78	28.33	73.12	42.93	61.80	35.70	33.00	11.39	37.51
OwlCap-7B	44.56	53.78	41.56	38.33	67.84	44.00	58.90	45.10	37.01	21.77	52.58
Qwen3-VL-30B-A3B-Instruct	52.04	61.00	51.33	43.78	72.48	45.87	62.50	51.32	41.02	21.50	60.05
Qwen3.6-35B-A3B	61.15	73.11	65.67	44.67	75.04	49.87	65.60	53.60	45.02	21.90	61.72
Qwen3-VL-235B-A22B-Instruct	53.81	61.33	53.56	46.56	73.12	48.53	63.90	47.80	43.18	**23.05**	61.26
CuRe	**65.44**	**73.67**	**66.33**	**56.33**	**77.44**	**51.73**	**67.80**	**54.43**	**47.86**	22.91	**63.56**

VideoMME, PerceptionTest for broad real-world video understanding covering diverse domains, video lengths, and perception-to-reasoning question types (Fu et al., 2025; Patraucean et al., 2023).

Results and analysis. As shown in Table 2, replacing the original Stage 2 captions with CuRe-generated captions consistently improves downstream video-language learning across three caption corpora. Averaged over all evaluation benchmarks, CuRe brings gains of 2.18, 1.75, and 2.04 points on Molmo2-Cap, Tarsier2-Recap, and LLaVA-Video-Caption, respectively. The gains extend across benchmarks: CuRe improves on the original captions in 25 out of 27 reported metrics, with the only regressions on MMVU for Molmo2-Cap and MVBench for LLaVA-Video-Caption.

Since the Stage 3 VideoQA mixture and the remaining training recipe are fixed within each paired comparison, the observed improvements isolate the contribution of Stage 2 caption quality. The consistent gains across three corpora with different source videos, scales, and annotating models further indicate that CuRe provides stronger video-language supervision for downstream VLM pretraining, with benefits extending beyond the scores assigned by a particular caption evaluator.

4.4 Caption-to-QA Prism Evaluation

We further adopt the Prism framework (Qiao et al., 2024) to evaluate captions as video proxies for video question answering (VQA). This tests whether CuRe-generated captions preserve enough task-relevant information for an answerer to address diverse questions without access to the video.

Settings. Prism evaluates caption utility via a two-stage VQA pipeline: each captioner first generates a video caption, and a fixed QA model, Seed 2.0 Pro, answers benchmark questions using only that caption. Official benchmark evaluators compute the final scores, measuring how much task-relevant information each caption exposes to the same answerer. We use the same set of benchmarks as introduced in Section 4.3 to evaluate the video caption quality on comprehensive VQA tasks.

Results and analysis. As shown in Table 3, CuRe achieves the strongest caption-to-QA performance among the evaluated open-source captioners. Its average score across the six benchmarks is 53.67, exceeding the same-sized Qwen3-VL-30B-A3B-Instruct by 5.60 points and the much larger

Table 4: Training-strategy comparison and claim-level reward ablation. The reward baselines differ in Reference-matching reward, Holistic reward, and CuRe. All GRPO variants share the SFT warm start, training videos, and GRPO budget. **Bold** indicates the best result for each metric.

Method	Training	DREAM-1K			VCapsBench		
		P↑	R↑	F1↑	AR↑	IR↓	CR↑
Training strategies							
SFT	SFT only	31.37	38.75	34.67	64.48	16.81	77.51
Continued SFT	SFT only	32.00	39.18	35.23	64.98	17.48	**78.76**
Reference-matching	SFT + GRPO	45.48	30.23	36.32	47.53	16.83	57.15
Holistic reward	SFT + GRPO	36.53	38.35	37.42	65.18	15.65	77.27
Claim-level reward components							
Base P/R	SFT + GRPO	42.03	35.99	38.78	65.08	16.59	78.03
Precision-only	SFT + GRPO	43.09	34.59	38.38	66.03	15.08	77.76
Recall-only	SFT + GRPO	36.35	**39.93**	38.06	65.69	16.30	78.48
CuRe	SFT + GRPO	**48.21**	36.26	**41.39**	**67.77**	**13.78**	78.60

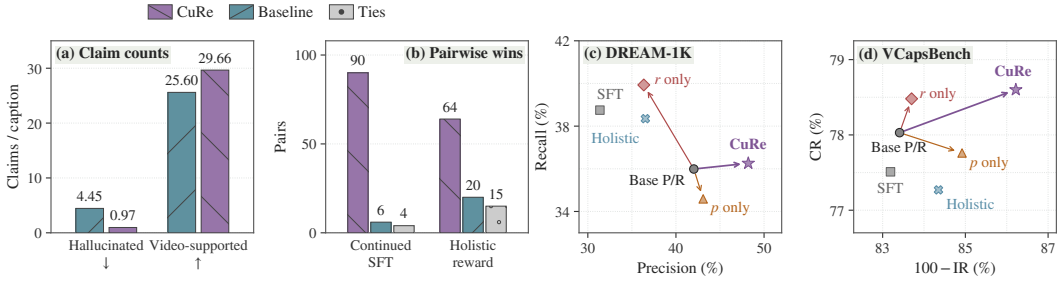


Figure 3: Human evaluation on 100 held-out videos: **(a)** mean hallucinated and video-supported claims per caption, against Continued SFT; **(b)** CuRe wins, baseline wins, and ties when jointly judging factual accuracy and semantic coverage in pairwise comparisons. Claim-level reward variants in **(c)** DREAM-1K precision–recall space and **(d)** VCapsBench $(100-\mathrm{IR})$–CR space, with SFT and Holistic reward for comparison. Arrows start from Base P/R. The labels p only and r only denote precision-only and recall-only training.

Qwen3-VL-235B-A22B-Instruct by 4.83 points. It also surpasses Qwen3.6-35B-A3B, the strongest open-source baseline on average, by 2.17 points. CuRe ranks first among open-source captioners on 10 out of 11 reported metrics. With the QA model held fixed, these results suggest that CuRe captions provide more task-relevant information for answering questions about the video content.

The improvement is particularly pronounced on long-form video understanding. Capturing key information for QA can be harder in long videos, where relevant details may be dispersed across an extended temporal span. On VideoMME Long, CuRe scores 56.33, outperforming Qwen3-VL-235B-A22B-Instruct by 9.77 points and Qwen3.6-35B-A3B by 11.66 points. It also comes within 0.89 points of the proprietary Seed 2.0 Pro captioner. This competitive performance suggests that CuRe captions can capture the key information needed to answer questions about long videos.

TOMATO remains a limitation, with CuRe scoring 22.91, slightly below Qwen3-VL-235B-A22B-Instruct at 23.05 and well behind the proprietary captioners, whose scores range from 35.31 to 35.58. This gap leaves substantial room for improvement in caption-based temporal reasoning. Nevertheless, the gains on the other benchmarks support our reward objective of balancing factual accuracy and semantic coverage to produce faithful and informative captions for downstream QA.

4.5 Ablation Study

The three settings above show that CuRe improves caption quality, downstream supervision, and caption-to-QA utility. We now analyze which components of the reward design drive these gains. We compare CuRe with baseline methods on DREAM-1K and VCapsBench and further ablate its reward components (Table 4). SFT is the shared warm start, and all GRPO variants share this initialization and training budget. Appendix C.3 details the shared training and evaluation settings.

Claim-level rewards reduce factual errors. Continued SFT extends the SFT warm start for one epoch on the same 8K videos as GRPO. Holistic reward uses a judge to assign a single overall score

to each caption based on the video and reference caption. Table 4 shows that the additional SFT epoch improves DREAM-1K recall and VCapsBench CR, yet raises IR from 16.81 to 17.48. Further reference-caption supervision therefore expands coverage without improving factual consistency. Holistic reward brings IR down to 15.65, but CuRe achieves 13.78 and surpasses this baseline in DREAM-1K precision and F1, as well as all three VCapsBench metrics. CuRe also maintains nearly the same coverage as Continued SFT, with CR scores of 78.60 and 78.76, respectively. The human evaluation in Figure 3(a,b) corroborates these gains: across 100 held-out videos, CuRe produces 0.97 hallucinated claims per caption compared with 4.45 for Continued SFT, while generating more video-supported claims. Annotators also prefer CuRe over both baselines when jointly assessing factual accuracy and semantic coverage. The annotation protocol is detailed in Appendix C.6.

Visual verification extends coverage beyond reference matching. Reference-matching reward scores captions against the reference text alone. Its DREAM-1K precision is high (45.48). Yet its recall (30.23) and VCapsBench CR (57.15) are the lowest in Table 4. This pattern suggests conservative captioning: the model favors details likely to match the reference and leaves more content undescribed. Figure 4 provides further support for this interpretation. Under Reference-matching reward, response length falls below 200 tokens during training and ends well below its initial level. By verifying claims against the video, CuRe can reward supported details absent from the reference. It raises DREAM-1K recall to 36.26 and VCapsBench CR to 78.60, while also improving precision. Holistic reward also uses video evidence and improves both coverage metrics over Reference-matching. Its responses, however, grow to roughly 500 tokens, whereas CuRe maintains a steadier length of around 390 tokens late in training, while retaining comparable coverage.

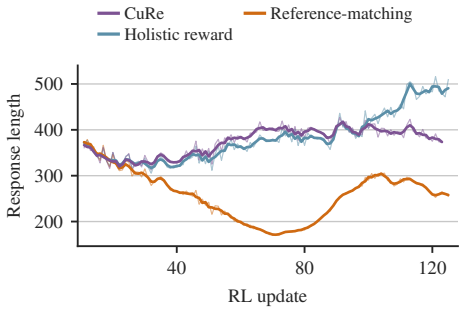


Figure 4: Mean response length per RL update. Faint traces are per-update values, bold lines a three-update mean.

Precision and recall rewards address complementary failure modes. Base P/R combines the basic claim-level scores without CuRe calibration. Relative to this control, Precision-only improves DREAM-1K precision and reduces VCapsBench IR. However, both recall and CR decline, suggesting that factual selectivity alone does not preserve coverage. Recall-only increases recall and CR, but DREAM-1K precision falls from 42.03 to 36.35. Figure 3(c,d) illustrates these opposing shifts in precision and recall. Combining the two terms gives CuRe the highest F1 of 41.39, compared with 38.38 for Precision-only and 38.06 for Recall-only (Table 4). On VCapsBench, CuRe also achieves the lowest IR and highest CR. These results show that joint optimization better balances factual accuracy and semantic coverage than either reward term alone.

We further examine category weighting and reward-model size in Appendix C.4. Uniform weighting improves factual selectivity over SFT but sacrifices semantic coverage. The proposed priors recover much of this lost coverage while further improving factual accuracy, supporting their role in balancing these objectives. Smaller reward models retain much of the recall and coverage achieved by the larger models. However, they yield lower precision and more factual inconsistencies than the larger models, indicating that factual accuracy remains sensitive to reward-model capacity.

5 Conclusion

We presented CuRe, a claim-level rubric reward framework for video captioning under incomplete reference supervision. By combining visual verification with reference-anchored credit weighting, CuRe prioritizes salient content, rewards additional video-supported details, and assigns zero credit to unsupported claims. Experiments show gains in factual accuracy with competitive semantic coverage, consistent improvements in downstream VLM pretraining, and stronger overall caption-to-QA performance than the evaluated open-source baselines. These findings support rewarding verified visual content beyond what a single reference caption describes. Our current study uses one policy backbone and a single reference-caption source. Future work will examine how well the reward generalizes across reference-caption sources, policy backbones, and video domains.

AI USE STATEMENT

In this work, we used generative AI tools to generate synthetic caption annotations, assist with method implementation and translation, clean datasets, and support qualitative analysis and interpretation of results. The use of generative models within our captioning, reward, and evaluation pipelines is described in the methodology, experiments, and appendix. We did not use generative AI tools to develop the conceptual framework, formulate mathematical claims, propose or refine research hypotheses, or design the research methodology or experiments; proof-related tasks are not applicable to this work. Additionally, we used generative AI tools for manuscript drafting and editing, literature identification, figure preparation, and visualization and typesetting code. We have reviewed all AI-assisted work, including checking code implementations, literature citations, and the numerical accuracy of figures and tables. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

We acknowledge the responsibilities set out in the ICLR Code of Ethics. This work uses existing video datasets, with data sources and privacy considerations documented in Appendix G and the human evaluation protocol in Appendix C.6. Although CuRe aims to reduce factual errors, generated captions may still contain hallucinations or reflect biases in the underlying data and models. Their use in privacy-sensitive or high-stakes applications requires additional review and human oversight.

REPRODUCIBILITY STATEMENT

To support reproducibility, Section 3 details the CuRe reward formulation. Appendices A and B document the training configuration, reward runtime, and category priors. Evaluation protocols and ablation settings are provided in Appendix C, with downstream VLM pretraining described in Appendix D. Appendix F provides prompts for claim decomposition, video verification, semantic matching, and caption-to-QA evaluation. Data sources are described in Appendix G.

REFERENCES

Daechul Ahn, Yura Choi, Youngjae Yu, Dongyeop Kang, and Jonghyun Choi. Tuning large multimodal models for videos using reinforcement learning from ai feedback. In *ACL 2024*, 2024. URL https://aclanthology.org/2024.acl-long.53/.

Kushal Arora, Layla El Asri, Hareesh Bahuleyan, and Jackie Chi Kit Cheung. Why exposure bias matters: An imitation learning perspective of error accumulation in language generation. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 700–710, 2022.

Anas Awadalla, Le Xue, Manli Shu, An Yan, Jun Wang, Senthil Purushwalkam, Sheng Shen, Hannah Lee, Oscar Lo, Jae Sung Park, Etash Guha, Silvio Savarese, Ludwig Schmidt, Yejin Choi, Caiming Xiong, and Ran Xu. BLIP3-KALE: Knowledge augmented large-scale dense captions, 2024. URL https://arxiv.org/abs/2411.07461.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report, 2025. URL https://arxiv.org/abs/2502.13923.

ByteDance Seed. Seed2.0 model card: Towards intelligence frontier for real-world complexity. https://lf3-static.bytednsdoc.com/obj/eden-cn/lapzild-tss/ljhwZthlaukjlkulzlp/seed2/0214/Seed2.0%20Model%20Card.pdf, 2026. Technical report / model card, accessed 2026-04-15.

David Chen and William Dolan. Collecting highly parallel data for paraphrase evaluation. In Dekang Lin, Yuji Matsumoto, and Rada Mihalcea (eds.), *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 190–200, Portland, Oregon, USA, June 2011. Association for Computational Linguistics. URL https://aclanthology.org/P11-1020/.

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, Li Yuan, Yu Qiao, Dahua Lin, Feng Zhao, and Jiaqi Wang. ShareGPT4Video: Improving video understanding and generation with better captions. 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. 19472–19495. Curran Associates, Inc., 2024. doi: 10.52202/079017-0614. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/22a7476e4fd36818777c47e666f61a41-Paper-Datasets_and_Benchmarks_Track.pdf.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models withăbetter captions. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (eds.), *Computer Vision – ECCV 2024*, pp. 370–387, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-031-72643-9.

Xiusi Chen, Gaotang Li, Ziqi Wang, Bowen Jin, Cheng Qian, Yu Wang, Hongru Wang, Yu Zhang, Denghui Zhang, Tong Zhang, Hanghang Tong, and Heng Ji. RM-R1: Reward modeling as reasoning. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 88313–88342, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/8e3b8de251afd887fb4589c1e3a3c793-Paper-Conference.pdf.

Jang Hyun Cho, Andrea Madotto, Effrosyni Mavroudi, Triantafyllos Afouras, Tushar Nagarajan, Muhammad Maaz, Yale Song, Tengyu Ma, Shuming Hu, Suyog Jain, Miguel Martin, Huiyu Wang, Hanoona Bangalath, Peize Sun, Po-Yao Huang, Daniel Bolya, Nikhila Ravi, Shashank Jain, Tammy Stark, Seungwhan Moon, Babak Damavandi, Vivian Lee, Andrew Westbury, Salman Khan, Philipp Kraehenbuehl, Piotr Dollar, Lorenzo Torresani, Kristen Grauman, and Christoph Feichtenhofer. Perceptionlm: Open-access data and models for detailed visual understanding. In D. Belgrave, C. Zhang, H. Lin,

R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 23475–23537. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/21f1c5bbf2519321c1bee9bfa9edcd46-Paper-Conference.pdf.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Mohammadreza Salehi, Rohun Tripathi, Sangho Lee, Zhongzheng Ren, Chris Dongjoo Kim, Yinuo Yang, Vincent Shao, Yue Yang, Weikai Huang, Ziqi Gao, Taira Anderson, Jianrui Zhang, Jitesh Jain, George Stoica, Winson Han, Ali Farhadi, and Ranjay Krishna. Molmo2: Open weights and data for vision-language models with video understanding and grounding, 2026. URL https://arxiv.org/abs/2601.10611.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, Ran He, and Xing Sun. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118, June 2025.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean Hendryx. Rubrics as rewards: Reinforcement learning beyond verifiable domains. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 127924–127945, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/cfd7bee7a651ee9af525098ef67a9e45-Paper-Conference.pdf.

Wenyi Hong, Yean Cheng, Zhuoyi Yang, Weihan Wang, Lefan Wang, Xiaotao Gu, Shiyu Huang, Yuxiao Dong, and Jie Tang. Motionbench: Benchmarking and improving fine-grained video motion understanding for vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8450–8460, June 2025.

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

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22195–22206, June 2024.

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. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 17417–17437, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.791. URL https://aclanthology.org/2026.acl-long.791/.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-ChatGPT: Towards detailed video understanding via large vision and language models. 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. 12585–12602, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.679. URL https://aclanthology.org/2024.acl-long.679/.

Desen Meng, Rui Huang, Zhilin Dai, Xinhao Li, Yifan Xu, Jun Zhang, Zhenpeng Huang, Meng Zhang, Lingshu Zhang, Yi Liu, and Limin Wang. Videocap-r1: Enhancing mllms for video captioning via structured thinking. *arXiv preprint*, 2025. URL https://arxiv.org/abs/2506.01725.

Gabriel Oliveira dos Santos, Esther Luna Colombini, and Sandra Avila. CIDEr-R: Robust consensus-based image description evaluation. In Wei Xu, Alan Ritter, Tim Baldwin, and Afshin Rahimi (eds.), *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pp. 351–360, Online, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.wnut-1.39. URL https://aclanthology.org/2021.wnut-1.39/.

Ramakanth Pasunuru and Mohit Bansal. Reinforced video captioning with entailment rewards. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel (eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 979–985, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1103. URL https://aclanthology.org/D17-1103/.

Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, joseph heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alexandre Fréchette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon Osindero, Dima Damen, Andrew Zisserman, and Joao Carreira. Perception test: A diagnostic benchmark for multimodal video models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 42748–42761. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/8540fba4abdc7f9f7a7b1cc6cd60e409-Paper-Datasets_and_Benchmarks.pdf.

Sundar Pichai, Demis Hassabis, and Koray Kavukcuoglu. A new era of intelligence with Gemini 3, November 2025. URL https://blog.google/products/gemini/gemini-3/.

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.

Qwen Team. Qwen3-VL technical report, 2025. URL https://arxiv.org/abs/2511.21631. Technical report; not peer-reviewed.

Ziyao Shangguan, Chuhan Li, Yuxuan Ding, Yanan Zheng, Yilun Zhao, Tesca Fitzgerald, and Arman Cohan. TOMATO: Assessing visual temporal reasoning capabilities in multimodal foundation models. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 7593–7734, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/16ba99f25a235f1100a4014d71d34ad8-Paper-Conference.pdf.

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.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. HybridFlow: A flexible and efficient RLHF framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, EuroSys '25, pp. 1279–1297, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400711961. doi: 10.1145/3689031.3696075. URL https://doi.org/10.1145/3689031.3696075.

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.

Mingyang Song and Mao Zheng. A survey of on-policy distillation for large language models. *arXiv preprint arXiv:2604.00626*, 2026.

Jiyang Tang, Hengyi Li, Yifan Du, and Wayne Xin Zhao. Avc-dpo: Aligned video captioning via direct preference optimization. *arXiv preprint arXiv:2507.01492*, 2025.

Jiawei Wang, Liping Yuan, Yuchen Zhang, and Haomiao Sun. Tarsier: Recipes for training and evaluating large video description models, 2024. URL https://arxiv.org/abs/2407.00634.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, Guanzhou Chen, Zichen Ding, Changyao Tian, Zhenyu Wu, Jingjing Xie, Zehao Li, Bowen Yang, Yuchen Duan, Xuehui Wang, Zhi Hou, Haoran Hao, Tianyi Zhang, Songze Li, Xiangyu Zhao, Haodong Duan, Nianchen Deng, Bin Fu, Yinan He, Yi Wang, Conghui He, Botian Shi, Junjun He, Yingtong Xiong, Han Lv, Lijun Wu, Wenqi Shao, Kaipeng Zhang, Huipeng Deng, Biqing Qi, Jiaye Ge, Qipeng Guo, Wenwei Zhang, Songyang Zhang, Maosong Cao, Junyao Lin, Kexian Tang, Jianfei Gao, Haian Huang, Yuzhe Gu, Chengqi Lyu, Huanze Tang, Rui Wang, Haijun Lv, Wanli Ouyang, Limin Wang, Min Dou, Xizhou Zhu, Tong Lu, Dahua Lin, Jifeng Dai, Weijie Su, Bowen Zhou, Kai Chen, Yu Qiao, Wenhai Wang, and Gen Luo. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025. URL https://arxiv.org/abs/2508.18265.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4581–4591, 2019.

Lipeng Xie, Sen Huang, Zhuo Zhang, Anni Zou, Yunpeng Zhai, Dingchao Ren, Kezun Zhang, Haoyuan Hu, Boyin Liu, Haoran Chen, Zhaoyang Liu, and Bolin Ding. Auto-rubric: Learning from implicit weights to explicit rubrics for reward modeling, 2026. URL https://arxiv.org/abs/2510.17314.

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 *ICLR 2026*, 2026. URL https://openreview.net/forum?id=JLelnhqXaC.

Tianwei Xiong, Yuqing Wang, Daquan Zhou, Zhijie Lin, Jiashi Feng, and Xihui Liu. Lvd-2m: A long-take video dataset with temporally dense captions. 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. 16623–16644. Curran Associates, Inc., 2024. doi: 10.52202/079017-0530. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/1df493ec1c2530c038d94d7300b5b368-Paper-Datasets_and_Benchmarks_Track.pdf.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5288–5296, 2016.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Dongjie Yang, Suyuan Huang, Chengqiang Lu, Xiaodong Han, Haoxin Zhang, Yan Gao, Yao Hu, and Hai Zhao. Vript: A video is worth thousands of words. 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. 57240–57261. Curran Associates, Inc., 2024. doi: 10.52202/079017-1824. URL https://proceedings.neurips.cc/paper_files/

paper/2024/file/6903a5aaece71b76623245fc6e32f01b-Paper-Datasets_and_Benchmarks_Track.pdf.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh Chawla, and Xiangliang Zhang. Justice or prejudice? quantifying biases in llm-as-a-judge. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 102351–102390, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/fdca08d371e4b6c031397909e20043bd-Paper-Conference.pdf.

Liping Yuan, Jiawei Wang, Haomiao Sun, Yuchen Zhang, and Yuan Lin. Tarsier2: Advancing large vision-language models from detailed video description to comprehensive video understanding, 2025. URL https://arxiv.org/abs/2501.07888.

Jiacheng Zhang, Yang Jiao, Shaoxiang Chen, Na Zhao, Zhiyu Tan, Hao Li, Xingjun Ma, and Jingjing Chen. Eventhallusion: Diagnosing event hallucinations in video llms, 2025. URL https://arxiv.org/abs/2409.16597.

Shi-Xue Zhang, Hongfa Wang, Duojun Huang, Xin Li, Xiaobin Zhu, and Xu-Cheng Yin. Vcaps-bench: A large-scale fine-grained benchmark for video caption quality evaluation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(15):12726–12734, Mar. 2026. doi: 10.1609/aaai.v40i15.38269. URL https://ojs.aaai.org/index.php/AAAI/article/view/38269.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. LLaVA-video: Video instruction tuning with synthetic data. *arXiv preprint*, 2024. URL https://arxiv.org/abs/2410.02713.

Yilun Zhao, Haowei Zhang, Lujing Xie, Tongyan Hu, Guo Gan, Yitao Long, Zhiyuan Hu, Weiyuan Chen, Chuhan Li, Zhijian Xu, Chengye Wang, Ziyao Shangguan, Zhenwen Liang, Yixin Liu, Chen Zhao, and Arman Cohan. Mmvu: Measuring expert-level multi-discipline video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8475–8489, June 2025.

Chunlin Zhong, Qiuxia Hou, Zhangjun Zhou, Yanhao Zhang, Shuang Hao, Haonan Lu, He Tang, and Xiang Bai. Owlcap: Harmonizing motion-detail for video captioning via hmd-270k and caption set equivalence reward. *arXiv preprint*, 2025. URL https://arxiv.org/abs/2508.18634.

APPENDIX OUTLINE

This appendix provides supplementary details to the main paper, organized as follows:

- (A) **Post-Training and Reward Runtime**
- (B) **Claim Schema and Category Priors**
- (C) **Evaluation and Reporting Protocols**
  - (C.1) Caption Quality Evaluation
  - (C.2) Prism Evaluation
  - (C.3) Reward Granularity and Reward-Component Ablation
  - (C.4) Category Weights and Reward-Model Size
  - (C.5) Reward-Module Quality
  - (C.6) Human Evaluation Protocol
  - (C.7) Reference Coverage Analysis
  - (C.8) Granularity of Verification and Matching
  - (C.9) Caption Length and Hallucination Density
- (D) **CuRe-Annotated Caption for VLM Pretraining**
- (E) **Qualitative Analysis**
- (F) **Prompt Templates**
  - (F.1) Reward Training Prompts
  - (F.2) Evaluation Prompts
- (G) **Data Privacy and Content Screening**

A POST-TRAINING AND REWARD RUNTIME

We first warm-start the captioner with SFT and then optimize it with GRPO using CuRe (§4.1). Table 5 reports the training configuration, and Table 6 specifies the reward runtime.

Table 5: Policy post-training configuration for the main CuRe model.

Stage	Parameter	Value
Supervised Fine-Tuning (SFT)		
	Data	78,144 Gemini-3.0-Pro distilled video caption samples
	Optimization	1 epoch; peak LR 4×10^{-6}; cosine decay; AdamW (β_1=0.9, β_2=0.999)
	Video budget	2 fps; max 512 frames; dynamic resolution (max 1,024 visual tokens per video)
GRPO Reinforcement Learning		
	Data	Disjoint 8,000-video subset (no overlap with SFT or evaluation)
	Rollout group size	$G = 8$ captions per video
	Rollout sampling	Temperature 1.0; top-p = 1.0; max 8,192 response tokens
	Actor LR / schedule	$2\times10^{-6} \rightarrow 2\times10^{-7}$; cosine decay; 5 warmup steps
	Batch size	Train 64, mini-batch 64, per-GPU micro-batch 1
	Sequence limits	24,576 prompt tokens; 8,192 response tokens
	KL control	Low-variance KL loss; $\beta = 0.005$; not added to reward
	Video budget	2 fps; max 256 frames; dynamic resolution
	Training duration	125 steps; ≈62.5 hours on 32×H20 (≈2,000 GPU-hours)
Infrastructure		
	Hardware	32 NVIDIA H20 GPUs
	Parallelism (actor)	TP=4, CP=2, EP=4, PP=1 (Megatron)
	Parallelism (rollout)	TP=4 (vLLM); GPU memory utilization 0.55
	Framework	verl 0.7.0 (Sheng et al., 2025); vLLM 0.11.0 (Kwon et al., 2023)
	Software	PyTorch 2.8.0; Transformers 4.57.3; Python 3.12
Data Integrity		
	Leakage control	SFT and GRPO videos are excluded from all evaluation benchmarks

Reward computation. For each GRPO rollout, CuRe decomposes the caption into category-typed claims, verifies them against the video, matches supported claims one-to-one to precomputed reference claims in the same category, and combines the resulting precision and recall scores with a length-control penalty to obtain the reward. Figures 10, 11, and 12 provide the decomposition, video verification, and semantic matching prompts. We accept a match when the matcher returns a valid reference-claim index with confidence at least 0.7. This cutoff follows the prompt's scoring rubric: scores of 0.7–0.8 denote the same core fact despite nonessential detail differences, while lower scores indicate missing distinguishing attributes or different facts. Accepted matches receive full credit, while video-supported claims without an accepted match retain residual credit $w_u = 0.5$, with their recall contribution capped as in Eq. 6.

Length-control regularizer. At reward time, we compute the generated-to-reference claim ratio $\rho = |\mathcal{C}|/|\mathcal{C}^\star|$ and subtract a bounded penalty

$$\Omega = \eta \cdot \min\left(1,\ \frac{\max(0,\ \rho - \tau)}{\kappa}\right), \tag{9}$$

where τ is the tolerance threshold, κ controls the penalty growth rate, and η is the maximum penalty. This saturated penalty suppresses verbose captions with excessively low-credit or redundant claims while maintaining stable optimization.

Table 6: Reward runtime configuration. The Qwen3-VL-235B roles use top-p = 1.0 (vLLM default). Gemini calls use Google Vertex API defaults for temperature and top-p. Video-consuming reward roles use the same 2-fps sampling and 256-frame cap as the policy rollout. Claim decomposition and semantic matching receive text only. Prompts are in Appendix F.

Role	Model	Temp	max_tok	Contract
Reference caption $y^\star$	Gemini-3.0-Pro	default	8,192	Generates the reference description for each training video
Rollout claims $\mathcal{C}$	Qwen3-VL-235B	0.3	4,096	Decomposes each policy caption into atomic visual claims
Reference claims $\mathcal{C}^\star$	Gemini-3.0-Pro	default	8,192	Decomposes each reference caption into atomic visual claims
Video verifier	Qwen3-VL-235B	0.3	4,096	Labels each claim as supported, conflicting, or uncertain
Semantic matcher	Qwen3-VL-235B	0.3	4,096	Returns one-to-one semantic alignments within each category
Reward Hyperparameters				
Matching confidence threshold		0.7		
Precision weight λ		0.7		(Eq. 8)
Residual credit w_u		0.5		(Eqs. 5–6)
Over-generation guard (η, τ, κ)		$(0.2,\ 1.5,\ 1.0)$		(Eq. 9)
Reward Video Processing				
Frame sampling		2 fps; max 256 frames; dynamic resolution from 256×256 to 768×768		

B Claim Schema and Category Priors

We organize caption claims into ten visual categories according to the type of information they convey. The schema covers entities and their properties, relations, actions and interactions, scene context, and other visual details. It provides a shared structure for claim decomposition, video verification, semantic matching, and reward aggregation. Grouping claims by information type also allows the reward to place different emphasis on different aspects of a video description.

Table 7: Claim categories and per-category priors used by CuRe.

Category	ω_k	Scope
Action & Interaction	1.8	Motion, behavior, interaction
Entity	1.2	People, animals, objects
Relation	1.2	Spatial or part-whole relations
Attribute	1.0	Color, size, shape, material, count
Environment	1.0	Scene, place, weather, background
Camera	0.8	Viewpoint, shot type, camera motion
Lighting	0.8	Light source, brightness, time of day
Emotion & Atmosphere	0.8	Mood conveyed by visible content
Text	0.8	Visible text, signs, subtitles
Visibility	0.8	Blur, occlusion, partial visibility
Default for unrecognized categories: 1.0		

We manually specify the category priors to emphasize event content in video captioning. Action & Interaction receives the highest weight to prioritize what happens in the video. Entity and Relation receive moderately higher weights to emphasize the participants and their spatial or part-whole relationships. Attribute and Environment retain the baseline weight, while the remaining categories receive lower positive weights so that additional visual details continue to contribute to the reward. Table 7 lists the corresponding weights. Compared with uniform weighting, this configuration improves both factual accuracy and semantic coverage in our experiments (Appendix C.4).

C Evaluation and Reporting Protocols

This section details the evaluation protocols corresponding to the experimental subsections in the main text. §C.1 covers caption quality evaluation (cf. §4.2), §C.2 covers Prism evaluation (cf. §4.4), and §C.3 covers the training and reward-design ablation (cf. §4.5). The downstream pretraining protocol is documented separately in §D.

C.1 Caption Quality Evaluation

Table 8 records the three direct caption evaluations used in Table 1. All entries are reported as percentages. Proprietary systems are included as reference upper bounds and are excluded from open-source bolding.

Table 8: Direct caption evaluation protocol for Table 1.

Benchmark	What is scored	Metrics	Notes
EventHallusion	Description-level caption quality on Entire, Interleave, and Misleading subsets	Entire, Interleave, Misleading, Overall	Higher is better for all scores
VCapsBench	Caption factual accuracy and semantic coverage via text-only QA	Accuracy Rate (AR), Inconsistency Rate (IR), Coverage Rate (CR)	Lower is better for IR; higher for AR and CR
DREAM-1K	Caption-level event precision and recall via AutoDQ extraction	Precision, Recall, F1	Evaluates factual accuracy and semantic coverage

Table 9 documents baseline grouping and result sourcing.

Table 9: Baseline grouping and result sourcing for evaluation. Reproduced results use official code and our evaluation pipeline; API-queried results use proprietary model APIs.

Group	Models	Source and notes
Proprietary references	Seed 2.0 Pro Gemini 3.0 Pro	**API-queried** Excluded from open-source bolding
Open-source captioners	Tarsier-34B Tarsier2-7B OwlCap-7B LLaVA-Video-7B	**Reproduced** Official code + default prompts
General-purpose VLMs	InternVL3.5-30B-A3B Qwen2.5-VL-72B-Instruct Qwen3-VL-30B-A3B-Instruct Qwen3-VL-235B-A22B-Instruct Qwen3.6-35B-A3B	**Reproduced** Default captioning prompt
CuRe	Qwen3-VL-30B-A3B-Instruct + CuRe post-training	**This work** Included in open-source set

C.2 Prism Evaluation

The Prism-style protocol evaluates whether a caption preserves enough information for a text-only QA model to answer benchmark questions. Table 10 separates the two stages.

The reported benchmarks are VideoMME, MMVU, MVBench, MotionBench, TOMATO, and PerceptionTest. VideoMME reports Overall plus Short, Medium, and Long splits; MMVU reports

Table 10: Two-stage Prism-style caption-to-QA protocol for Table 3.

Stage	Variable	Input and output
Stage 1 captioning	Captioner row in Table 3	The captioner receives the video and the caption prompt in Figure 15; it outputs a text caption
Stage 2 QA	Fixed QA model	Seed 2.0 Pro receives the generated caption and benchmark question. Multiple-choice inputs also include answer options (Figure 16); MMVU Open uses the free-text answering prompt (Figure 17)
Scoring	Fixed benchmark evaluator	Multiple-choice answers are checked against the answer key. MMVU Open answers are judged against reference answers using Figure 18

MCQ, Open, and Overall; the remaining benchmarks report Overall. MMVU Open uses caption-conditioned free-text answering followed by reference-based answer judging (Figures 17 and 18). Reference answers are provided only to the judge.

C.3 Reward Granularity and Reward-Component Ablation

Table 11 defines the rows in Table 4. All GRPO rows start from the same SFT initialization and use the same training videos, rollout and optimization configuration, and number of training steps; only the reward changes.

Table 11: Training controls and RL reward definitions for Table 4. Continued SFT matches the GRPO training data under the SFT objective. All GRPO rows use the same initialization, training videos, rollout and optimization configuration, and number of training steps; only the reward changes.

Row	Training	Final RL reward	Purpose
SFT	SFT only	–	Measures the warmstarted captioner before RL post-training
Continued SFT	SFT only	–	Data-matched SFT baseline using the same 8K videos and Gemini targets as GRPO
Reference-matching	SFT + GRPO	Text-only consistency and coverage, Eq. 10	Tests treating the reference as the complete reward target; the judge receives no video and performs no claim decomposition
Holistic reward	SFT + GRPO	Video- and reference-grounded whole-caption score	Tests response-level reward feedback without claim decomposition, claim-level verification, or claim matching
Base P/R	SFT + GRPO	$\lambda p^{\mathrm{base}} + (1-\lambda) r^{\mathrm{base}} - \Omega$	Tests the basic claim-level control without CuRe calibration
Precision-only	SFT + GRPO	$p^{\mathrm{CuRe}} - \Omega$	Isolates the precision-side selective-credit term
Recall-only	SFT + GRPO	$r^{\mathrm{CuRe}} - \Omega$	Isolates the recall-side bounded reference-target expansion term
Full CuRe	SFT + GRPO	$\lambda p^{\mathrm{CuRe}} + (1-\lambda) r^{\mathrm{CuRe}} - \Omega$	Combined precision-recall reward used by the main model

Shared claim-level reward controls. Base P/R, Precision-only, Recall-only, and full CuRe all subtract the same length-control penalty Ω from Eq. 9, with the same (η, τ, κ) settings. They also share the category priors and active-category aggregation in Eq. 7. For Base P/R, p^{base} and r^{base} aggregate the basic per-category scores in Eq. 4; full CuRe instead aggregates the calibrated scores in Eqs. 5 and 6. Both combined rewards use $\lambda = 0.7$. Precision-only and Recall-only retain the corresponding calibrated aggregate and the same penalty, isolating each reward component.

Continued SFT. Continued SFT starts from the same SFT warm start and continues SFT for one epoch on the same 8,000-video training set used by the GRPO runs. This forms the data-matched SFT baseline in Table 4.

Reference-matching reward. Reference-matching uses Qwen3-VL-235B-A22B-Instruct to compare the complete candidate caption with its Gemini reference caption using text alone. The policy receives the video as usual, but the reward judge receives neither video frames nor decomposed claims. The judge returns two scores in $[0, 1]$: consistency, measuring support for candidate content in the reference, and coverage, measuring how much reference content the candidate recovers. We combine them as

$$R_{\text{ref}} = 0.7\, s_{\text{consistency}} + 0.3\, s_{\text{coverage}}. \tag{10}$$

The judge scores consistency and coverage at the caption level. The prompt treats the reference as the complete evaluation target and withholds consistency credit from details absent from it. The resulting weighted score directly supplies the reward. The run uses the same SFT initialization, 8,000 training videos, rollout and optimization settings, and number of training steps as full CuRe. Figure 14 gives the exact English prompt.

Holistic reward. The Holistic reward baseline uses Qwen3-VL-235B-A22B-Instruct as a single whole-caption judge, matching the reward model used in the other GRPO rows of Table 4. For each rollout, it receives the video, the complete candidate caption, and the corresponding Gemini-3.0-Pro reference caption. Instead of decomposing the candidate caption into claims, the judge rates the caption as a whole along two criteria, factuality and informativeness, and returns a single overall score in $[0, 1]$ that serves as the scalar reward. The policy uses the same SFT initialization, 8,000 training videos, rollout and optimization configuration, and number of training steps as the other GRPO variants, with only the reward function changed. Figure 13 gives the full prompt.

C.4 Category Weights and Reward-Model Size

Table 12 addresses two questions: do the category priors improve caption quality, and how sensitive is CuRe to the size of its reward models? Each variant retains the claim-level pipeline and changes one setting. We repeat SFT and full CuRe from Table 4 for comparison.

Table 12: Category-weight and reward-model-size ablations on DREAM-1K and VCapsBench. SFT and full CuRe are repeated from Table 4. The ablations set all category priors to 1.0 or replace all three online 235B reward models with 30B models. **Bold** marks the best value in each metric.

Variant	DREAM-1K			VCapsBench		
	P↑	R↑	F1↑	AR↑	IR↓	CR↑
SFT	31.37	**38.75**	34.67	64.48	16.81	77.51
CuRe w/ uniform weights	44.33	27.32	33.80	56.87	15.21	67.07
CuRe w/ 30B judges	43.46	34.99	38.77	63.91	15.55	75.67
Full CuRe (235B judges)	**48.21**	36.26	**41.39**	**67.77**	**13.78**	**78.60**

Category priors improve caption quality. We retain the ten-category schema and set $\omega_k = 1.0$ for every category in Eq. 7, replacing the priors in Table 7. The precision–recall mixture $\lambda = 0.7$, residual credit $w_u = 0.5$, and remaining reward settings are unchanged. Using the proposed priors improves DREAM-1K recall by 8.94 points and F1 by 7.59 points over uniform weighting, while raising VCapsBench AR and CR by 10.90 and 11.53 points, respectively. Under uniform weighting, F1 (33.80) and CR (67.07) fall below SFT (34.67 and 77.51), despite higher precision and lower IR. The proposed category priors improve all six metrics over equal weighting, supporting their role in preserving semantic coverage while improving factual accuracy.

Sensitivity to reward-model size. To test whether smaller models can replace the 235B reward models, we use Qwen3-VL-30B-A3B-Instruct in all three online roles: rollout claim decomposition, video verification, and semantic matching. The policy model remains Qwen3-VL-30B-A3B, and the Gemini reference captions and precomputed reference claims are unchanged. The 30B variant reaches 38.77 DREAM-1K F1, 4.10 points above SFT and 2.62 below full CuRe. Its VCapsBench

AR, IR, and CR are 63.91, 15.55, and 75.67, respectively, with CR 2.93 points below full CuRe. It retains gains over SFT in DREAM-1K precision and F1 and in VCapsBench IR, but not in recall, AR, or CR. The pipeline therefore supports 30B judges, with lower caption quality than the 235B configuration on all six metrics. Appendix C.5 evaluates the individual reward modules to examine this dependence on model size.

C.5 Reward-Module Quality

We evaluate decomposition, verification, and matching separately to examine how each module's quality changes with model size. Qwen3-VL-235B-A22B-Instruct and Qwen3-VL-30B-A3B-Instruct process the same set of 100 videos and their captions. Gemini-3.1-Pro evaluates every output using the video sampled at 1 fps, with temperature set to 0. The 235B and 30B pipelines produce 3,298 and 3,534 claims, respectively. Matching is evaluated on their 3,029 and 3,391 claims labeled supported by the verifier.

Decomposition validity. We assess each extracted claim for atomicity and verifiability. An atomic claim expresses a single visual property or phenomenon that can be checked independently; a verifiable claim can be assessed against visual evidence. Decomposition validity is the proportion of extracted claims that Gemini judges to satisfy both criteria. This evaluates the form of the decomposition: a verifiable claim need not be true of the video, and visual support is assessed separately by the verification module.

Verification accuracy and false support. For each extracted claim, we compare the verifier's label with Gemini's label under the same three-way classification: *supported* when the claim is directly observable in the video, *conflicting* when it contradicts visible content, and *uncertain* when the evidence is insufficient to confirm or reject it. Verification accuracy is the proportion of claims receiving the same label from both judges. The false-support rate instead uses only claims labeled supported by the pipeline's verifier as its denominator, and counts those labeled conflicting by Gemini in its numerator. Gemini-uncertain claims are not included in that numerator.

Matching accuracy. We evaluate matching on each pipeline's verifier-supported claims. The matcher aligns these with reference claims in the same visual category under a one-to-one constraint; semantic equivalence concerns the same visual fact, rather than identical wording. Matching accuracy measures agreement between the reference-claim index predicted by the matcher and the index selected by Gemini, over the evaluated verifier-supported claims. Thus, verification accuracy measures agreement about video support, whereas matching accuracy measures agreement about reference alignment.

Table 13 and Figure 5 report the four metrics. The two pipelines generate different claim sets, so their claim-level evaluation populations differ despite using the same videos and captions.

Table 13: **Module evaluation scores.** Values are percentages. Δ is 30B minus 235B in percentage points; bold marks the better score.

Metric	235B	30B	Δ
Decomposition validity ↑	81.23	**86.11**	+4.88
Verification accuracy ↑	**81.93**	76.71	−5.22
False-support rate ↓	**6.17**	10.03	+3.86
Matching accuracy ↑	**87.32**	84.90	−2.42

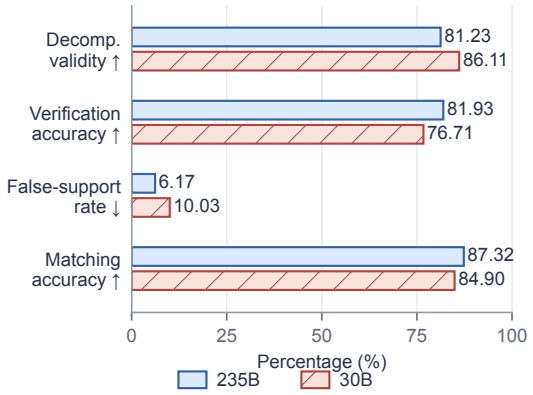


Figure 5: **Reward-module quality.** Qwen3-VL-235B versus 30B. Values are percentages; ↑/↓ indicate higher/lower is better.

Comparing judges on identical claims. We also fix the 235B decomposer's output and run both judges on the same 3,298 claims. In this comparison, the 30B verifier processes the claims in caption-level batches. Verification labels agree on 95.21% of claims, with Cohen's $\kappa = 0.580$. For matching, we use the same 3,029 claims marked supported by the 235B verifier and apply a confidence threshold of 0.7. The two models select the same reference index, including a no-match outcome, on 88.54% of these claims. Gemini-3.1-Pro adjudicates all 158 verification disagreements and 347 matching disagreements (Table 14). On verification disagreements, 235B is correct in 65.19% of cases. Matching disagreements are more evenly split, with 235B correct in 54.18% of cases and 30B in 45.82%.

Table 14: 235B and 30B judges evaluated on identical input claims from 100 captions. Agreement is measured over all claims for each task. Gemini-3.1-Pro determines the correct judge output on the disagreement subset. Percentages in the last two columns use that subset as the denominator.

Task	Claims	Agreement (%)	Disagreements	235B correct (%)	30B correct (%)
Verification	3,298	95.21	158	65.19	34.81
Matching	3,029	88.54	347	54.18	45.82

30B is a viable alternative for the text modules. Table 13 shows that 30B improves decomposition validity from 81.23% to 86.11%, while matching accuracy decreases from 87.32% to 84.90%. Both modules operate on text: decomposition extracts atomic claims from captions, and matching aligns supported claims with reference claims. The higher validity makes decomposition the clearest candidate for replacement with 30B. Matching is also a practical candidate, with a 2.42-point accuracy trade-off.

Video verification benefits most from the larger judge. Replacing 235B with 30B reduces verification accuracy from 81.93% to 76.71% and raises the false-support rate from 6.17% to 10.03%. Unlike the text modules, verification must establish whether each claim is supported by visual evidence. False support decisions admit visually conflicting claims into the set eligible for reward. This pattern is consistent with the GRPO results in Table 12: using 30B in all three roles lowers DREAM-1K precision from 48.21 to 43.46 and raises VCapsBench IR from 13.78 to 15.55. The 30B pipeline still provides useful training signals, improving DREAM-1K F1 over SFT from 34.67 to 38.77, but remains below the 235B configuration's 41.39. Taken together, the module evaluations and training results favor retaining 235B for video verification while prioritizing the text modules for replacement with 30B.

C.6 Human Evaluation Protocol

We evaluate CuRe, Continued SFT, and Holistic reward on 100 videos held out from RL training (Section 4.5). The models share the SFT initialization and 8K post-training videos. Six annotators work in three pairs. Both members independently assess each assigned sample after watching the video. Model identities are hidden, and caption order is randomized. Disagreements are reviewed by an annotator from another pair and resolved by majority vote.

For CuRe and Continued SFT, annotators count claims contradicted by the video (hallucinated claims) and claims supported by it (video-supported claims). Figure 3(a) reports mean counts per caption. In the pairwise task, annotators compare captions of the same video for factual accuracy and semantic coverage, allowing ties. Figure 3(b) reports wins, losses, and ties for CuRe across 100 valid pairs against Continued SFT and 99 against Holistic reward.

C.7 Reference Coverage Analysis

Reference matching leaves many verifier-supported claims without matching credit. We analyze 700 claims marked supported by the video verifier across 100 captions. Matching against the reference claim set $\mathcal{C}^\star$ covers 401 claims (57.29%), leaving 299 (42.71%) unmatched. Under a reward that credits only text-reference matches, these 299 claims would receive zero claim-level credit despite passing video verification.

We use Gemini-3.0-Pro to recheck all 299 unmatched claims directly against their videos. It identifies 288 correct claims (96.32%), 10 hallucinated claims (3.34%), and one uncertain claim (0.33%), as reported in Table 15.

CuRe assigns a discounted credit of $w_u = 0.5$ to each video-supported claim without a reference match. This rewards additional correct visual details, with an observed hallucination rate of 3.34% among the unmatched claims.

Table 15: Reference coverage of 700 claims marked supported by the video verifier across 100 captions, with Gemini-3.0-Pro verification of the 299 unmatched claims.

Reference matching (700 claims)		
Outcome	Count	Share (%)
Matched	401	57.29
Unmatched	299	42.71
Video-based recheck of unmatched claims (299 claims)		
Outcome	Count	Share (%)
Correct	288	96.32
Hallucinated	10	3.34
Uncertain	1	0.33

C.8 Granularity of Verification and Matching

Grouping claims by visual category improves both verification and matching. We evaluate verification and matching on separate caption corpora with independent claim sets. For each task, we compare four ways to organize judge calls: all claims in one batch, four coarse groups (Content, Interaction, Setting, and Cinematography), the ten categories used by CuRe, and one claim per call. All variants use Qwen3-VL-235B-A22B with Gemini-3.1-Pro labels for evaluation. Matching is one-to-one within each visual category in all four settings.

Table 16 reports verification Macro-F1, the unweighted mean of F1 for supported, conflicting, and uncertain labels, together with conflicting-claim F1 and matching accuracy. Verification uses 2,904 valid claims for the batch-all, coarse-group, and per-category variants, and 2,896 for per-claim calls. Matching uses 3,395 supported claims for each variant.

Table 16: Effect of claim grouping on Qwen3-VL-235B verification and matching, evaluated against Gemini-3.1-Pro labels. Quality scores are percentages. n_{verify} is the number of valid verification claims. Matching uses 3,395 claims in every row. Calls are approximate verification calls per caption. Bold marks the best quality score in each column.

Grouping	n_{verify}	Verification Macro-F1 ↑	Conflict F1 ↑	Matching accuracy ↑	Calls ↓
All claims	2,904	36.6	11.5	42.74	1
Four coarse groups	2,904	45.8	28.5	80.91	∼4
Ten categories (CuRe)	2,904	46.3	29.8	84.80	∼10
Per claim	2,896	**46.6**	**37.6**	**88.75**	∼35

Per-category calls raise verification Macro-F1 from 36.6% to 46.3% and conflicting-claim F1 from 11.5% to 29.8% over a single batch. Matching accuracy increases from 42.74% to 84.80%. Finer per-claim calls reach 46.6% Macro-F1 and 88.75% matching accuracy, using approximately 35 verification calls per caption. The ten-category grouping achieves nearly the same verification Macro-F1 with approximately ten calls, while improving matching accuracy by 3.89 points over four coarse groups. These results support category-based batching as an effective balance between judgment quality and call count.

C.9 Caption Length and Hallucination Density

CuRe produces longer captions than Continued SFT while reducing hallucinations per unit of text. We compare all three systems on 168 videos with valid captions from each system. We assess each claim against its video, labeling claims as hallucinated when contradicted by the video and as supported when explicitly supported by visual evidence. We estimate caption length using 1.3 times the word count for English and the character count for Chinese. Mean claim counts are divided by mean caption length and scaled to 100 estimated tokens in Table 17. Figure 6 plots mean length against hallucination density for the same three systems.

Table 17: **Caption statistics on 168 shared videos.** Mean length is in estimated tokens. Halluc. and Supported count claims contradicted and supported by the video, respectively. Each density is 100 times the mean claim count divided by mean caption length. Bold marks the best density in each column.

Method	Mean length	Halluc. density ↓	Supported density ↑
Continued SFT	455.5	0.98	5.26
Holistic reward	622.8	0.41	5.41
CuRe	567.7	**0.20**	**5.62**

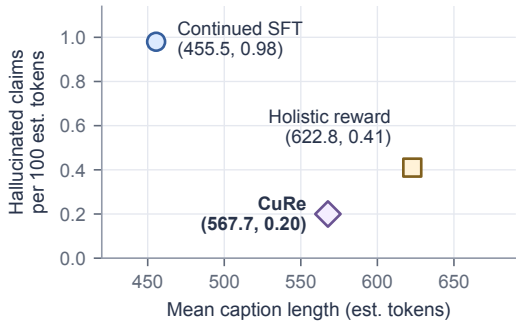


Figure 6: **Caption length and hallucination density.** Each point is a system mean over the same 168 videos. Lower hallucination density is better.

CuRe averages 567.7 estimated tokens per caption, compared with 455.5 for Continued SFT and 622.8 for Holistic reward. Its hallucination density is 0.20 per 100 estimated tokens, compared with 0.98 and 0.41 for the two baselines. Relative to Continued SFT, CuRe provides 24.6% more text with approximately 80% lower hallucination density. It also has lower hallucination density than Holistic reward while producing shorter captions. Its supported-claim density is 5.62 per 100 estimated tokens, compared with 5.26 for Continued SFT and 5.41 for Holistic reward, indicating more video-supported information per unit of text.

D CuRe-Annotated Caption for VLM Pretraining

This section gives the training details for the downstream pretraining utility study in Table 2. Following the ShareGPT4V (Chen et al., 2025; 2024) and CapRL (Xing et al., 2026) pretraining paradigm, we use three stages: initial visual-language alignment, caption-data further pretraining, and instruction SFT. The study is designed as a controlled data-quality test: within each paired row, the model architecture, Stage 1 alignment, Stage 3 SFT, and evaluation suite are fixed, and only the Stage 2 caption source changes.

Model architecture. The downstream learner combines the pretrained vision transformer (ViT) from Qwen3-VL-4B with the Qwen3-4B Base language model through a randomly initialized multimodal projector (Kaiming normal initialization). The language backbone is loaded from the base LLM checkpoint. Stage 1 trains the projector to align the pretrained visual features with the language model while keeping both the ViT and the LLM frozen.

Training setting. Stage 1 performs initial alignment on BLIP-3-Kale image-caption data. Only the projector is trainable while the ViT and LLM are frozen. Stage 2 performs further pretraining on video-caption data with all modules unfrozen. For each Stage 2 corpus, we train on either the original captions or on CuRe-generated recaptions. The reported paired corpora are Molmo2-Cap, Tarsier2-Recap, and LLaVA-Video-Caption. Stage 3 performs VideoQA instruction SFT starting from the corresponding Stage 2 checkpoint, using the same merged VideoQA data (Molmo2-AskModelAnything 20K + LLaVA-Video-178K Stage-3 80K) for every paired comparison.

Controlled comparisons. Table 18 lists the hyperparameters for the three stages. The paired comparisons in Table 2 keep these settings fixed within each corpus pair. We report MMVU, MVBench,

MotionBench, PerceptionTest, TOMATO, and VideoMME accuracies. When computing the average gain, we average benchmark-level overall columns; the VideoMME duration splits are diagnostic and are not counted as separate benchmarks.

Table 18: Hyperparameters for the downstream pretraining utility study. Stage 2 is the only stage whose caption source changes across paired rows in Table 2.

Stage	Trainable modules	Data	Optimization	Length / video budget
Initial Alignment	Projector / aligner only; ViT and LLM frozen	BLIP-3-Kale image-caption JSONL	1 epoch; LR 1×10^{-3}; min LR 1×10^{-4}; warmup 0.05; global / micro batch 512 / 64	Max length 8,192; image max token 1,024
Further Pretraining	ViT, LLM, and projector / aligner all unfrozen	One Stage 2 caption corpus per run; original captions or CuRe recaptions	1 epoch; LR 4×10^{-5}; min LR 2×10^{-6}; warmup 0.03; global / micro batch 64 / 1	Max length 64,000; video max token 1,024; 2 fps; at most 512 frames
VideoQA SFT	ViT, LLM, and projector / aligner all unfrozen	Molmo2-AskModelAnything 20K + LLaVA-Video-178K Stage-3 80K by default	1 epoch; LR 1×10^{-5}; min LR 2×10^{-6}; warmup 0.03; global / micro batch 128 / 1	Max length 64,000; same video budget as Stage 2

E QUALITATIVE ANALYSIS

Figures 7, 8, and 9 show three representative side-by-side caption comparisons on held-out DREAM-1K videos. Figure 7 compares CuRe against its same-size baseline (Qwen3-VL-30B-A3B-Instruct); Figures 8 and 9 compare CuRe against the much larger Qwen3-VL-235B-A22B-Instruct. Captions are excerpts from each model's output; "..." indicates omitted unrelated text. Red marks hallucinated or unsupported baseline content; green marks the core correct event; blue marks visually supported details that CuRe adds.

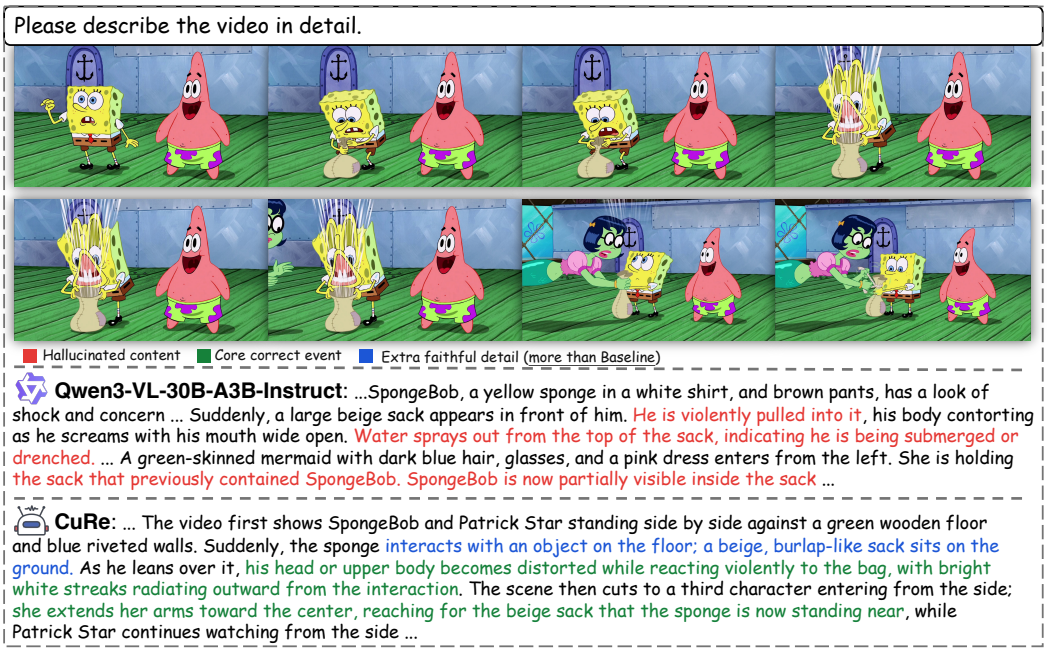


Figure 7: Qualitative comparison on a held-out DREAM-1K video. Qwen3-VL-30B-A3B-Instruct recognizes the sack interaction but adds unsupported causal details, such as water spraying out and SpongeBob being submerged or still inside the sack. CuRe keeps the shared event while adding visible details about the sack, impact streaks, and the later reaching action. Red marks hallucinated content, green marks the core correct event, and blue marks faithful details added by CuRe.

Figure 7 illustrates the balance between factual accuracy and semantic coverage in one held-out DREAM-1K video. Both captions find the central sack interaction, so the difference is not basic scene recognition. The baseline turns the cartoon impact lines around the sack into unsupported causal content: water sprays out, and SpongeBob is being submerged or drenched. It also treats the later sack as the same sack that previously contained SpongeBob. These details fit a plausible story, but they are not directly supported by the sampled frames.

CuRe describes the same event through visible relations. It names the burlap-like sack on the floor, the distorted reaction near the sack, and the later reaching action toward the sack. Rather than shortening the description, the caption changes the evidence type: inferred causes become verifiable visual claims. This matches the reward design in §3.2: unsupported claims receive no credit, while video-supported residual details can still receive bounded credit when they fall outside the reference caption. The case mirrors the pattern in Table 1 and Table 4: CuRe improves factual accuracy without suppressing informative detail.

Please describe the video in detail.



Hallucinated content Core correct event Extra faithful detail (more than Baseline)

Qwen3-VL-235B-A22B-Instruct: ... The video captures an exhilarating downhill race on a winding mountain road, filmed from the first-person perspective of a rider on a gravity-powered vehicle - likely a street luge or downhill longboard. ... Ahead, a group of competitors races down the same road, each on similar low-profile boards. They wear colorful racing suits and helmets ...

CuRe: ... This video presents a first-person POV from a rider in a downhill race on a paved mountain road. The rider's hands grip handlebars, and a large front wheel is partially visible ... Ahead of the POV rider, a group of other competitors is navigating the descent. These participants are riding specialized low-profile vehicles, with riders adjusting their legs through the bend. The camera wearer closes the gap, overtakes from the left, and prepares for the next left turn.

Figure 8: Caption comparison on a held-out DREAM-1K downhill race video. Red marks hallucinated content, green marks the core correct event, and blue marks faithful details added by CuRe.

Please describe the video in detail.



Hallucinated content Core correct event Extra faithful detail (more than Baseline)

Qwen3-VL-235B-A22B-Instruct: ... In a lush, sun-dappled forest clearing, two animated creatures stand on a dirt path ... The larger creature is a massive, muscular white rabbit ... Beside it, a much smaller, dark gray mouse holds a shiny acorn in its tiny paws. ... The mouse, undeterred, continues to hold the acorn and looks around, contemplating the rabbit's words ...

CuRe: ... The video presents a 3D animated forest scene with a giant rabbit figure and a small dark-furred creature. The small creature holds a round reddish-brown nut and reacts to the rabbit's direction. As the sequence progresses, the large creature extends an arm or points, and the small creature moves toward the green bush. The bush is then revealed as a disguised wooden trap, turning the exchange into a directed trap setup rather than a simple forest conversation.

Figure 9: Qualitative comparison on a held-out DREAM-1K video of an animated forest scene. Red marks hallucinated content, green marks the core correct event, and blue marks faithful details added by CuRe.

F PROMPT TEMPLATES

This section lists the prompt contracts used in reward training and evaluation. Prompts are grouped into reward-training prompts (§F.1) used by the online reward runtime, and evaluation prompts (§F.2) used for direct caption and Prism-style evaluation.

F.1 REWARD TRAINING PROMPTS

The reward prompts reproduce the English templates used by the claim decomposer, video verifier, matcher, and the whole-caption reward judges. Placeholders such as `{caption}` denote values filled at runtime. Figure 10 shows the shared claim-schema decomposition prompt used for rollout claims and precomputed Gemini reference claims. Figure 11 shows video-only claim verification. Figure 12 shows the per-category one-to-one matcher. Figure 13 shows the whole-caption prompt used by the Holistic reward baseline, and Figure 14 gives the text-only Reference-matching prompt.

Claim Decomposition | Reward Prompt

```
You are a visual information analysis assistant. Decompose the following
video/image description into independent atomic visual facts.
### Rules
1. **Atomicity**: Each fact describes exactly one directly observable property or
phenomenon. Separate entity presence, quantity, individual attributes, spatial
relations, and actions. Camera movement, movement speed, viewpoint, and framing
are also separate facts.
2. **Explicit subjects**: Use original nouns (e.g., "the cat", "the red car").
Never use pronouns like "it" or "this". Use the full description on first mention,
then a shorter form.
3. **No redundancy**: Do not repeat the same attribute across facts.
4. **Temporal & dynamic**: Camera movements (pan, tilt, zoom), scene transitions,
and motion trajectories should each be an independent fact.
5. **Category constraint**: The category field must be one of the following enum
values:
[Entity, Attribute, Relation, Action & Interaction, Environment, Lighting, Camera,
Emotion & Atmosphere, Text, Visibility]
### Category Reference
| Enum Value | Scope |
|--------|-----|
| Entity | People, animals, objects appearing in the scene |
| Attribute | Color, size, shape, material, quantity of entities |
| Relation | Spatial relations (left of, in front of), part-of relations |
| Action & Interaction | Movements, behaviors, interactions between entities |
| Environment | Scene, location, weather, background elements |
| Lighting | Light sources, direction, brightness, time of day |
| Camera | Camera movement, angle changes, filming techniques |
| Emotion & Atmosphere | Mood, atmosphere conveyed by the visuals |
| Text | Visible text, signs, subtitles in the frame |
| Visibility | Occlusion, blur, image quality, partial visibility |
### Example
Description: An aerial video of a city at night. The camera slowly pushes
forward, overlooking the entire city. Tall buildings fill the frame with colorful
lights. Traffic flows on the roads, headlights forming light trails. A Ferris
wheel with blue lights stands on the right side. The overall tone is warm,
conveying a bustling urban atmosphere.
Decomposition:
```json
[
{"category": "Environment", "fact": "The scene shows a city", "is_visual":
"yes"},
{"category": "Lighting", "fact": "The scene takes place at night", "is_visual":
"yes"},
{"category": "Camera", "fact": "The camera has an overhead viewpoint",
"is_visual": "yes"},
{"category": "Camera", "fact": "The camera pushes forward", "is_visual": "yes"},
{"category": "Camera", "fact": "The camera moves slowly", "is_visual": "yes"},
{"category": "Camera", "fact": "The camera view covers the entire city",
"is_visual": "yes"},
```
```

Claim Decomposition | Reward Prompt *(continued)*

```
{"category": "Entity", "fact": "Buildings are visible in the scene", "is_visual":
"yes"},
{"category": "Attribute", "fact": "There are multiple buildings", "is_visual":
"yes"},
{"category": "Attribute", "fact": "The buildings are tall", "is_visual": "yes"},
{"category": "Camera", "fact": "The buildings fill the frame", "is_visual":
"yes"},
{"category": "Lighting", "fact": "Lights illuminate the buildings", "is_visual":
"yes"},
{"category": "Attribute", "fact": "The building lights are colorful",
"is_visual": "yes"},
{"category": "Entity", "fact": "Traffic is visible in the scene", "is_visual":
"yes"},
{"category": "Relation", "fact": "The traffic is on the roads", "is_visual":
"yes"},
{"category": "Action & Interaction", "fact": "The traffic is moving",
"is_visual": "yes"},
{"category": "Lighting", "fact": "The headlights form light trails", "is_visual":
"yes"},
{"category": "Entity", "fact": "A Ferris wheel is visible in the scene",
"is_visual": "yes"},
{"category": "Relation", "fact": "The Ferris wheel is on the right side of the
frame", "is_visual": "yes"},
{"category": "Lighting", "fact": "Lights illuminate the Ferris wheel",
"is_visual": "yes"},
{"category": "Attribute", "fact": "The Ferris wheel lights are blue",
"is_visual": "yes"},
{"category": "Lighting", "fact": "The overall color tone of the scene is warm",
"is_visual": "yes"},
{"category": "Emotion & Atmosphere", "fact": "The scene conveys a bustling urban
atmosphere", "is_visual": "yes"}
]
```

```
### Description to Decompose
{caption}
Briefly analyze the description, then output the decomposition in the following
JSON format:
```json
[
{"category": "<enum value>", "fact": "<complete independent visual fact>",
"is_visual": "<yes or no>"},
...
]
```
Ensure valid JSON format. All field values must be strings. Do not include
comments in JSON.
```

Figure 10: Full prompt template for claim-schema decomposition in reward training.

Video Verification | Reward Prompt

```
Watch the video/image carefully and judge whether each of the following visual
facts is accurate.
### Judgment Criteria
- **support**: The fact can be directly observed in the video/image. For videos,
appearing at any point in time counts as support.
- **conflict**: The fact explicitly contradicts the visible content (e.g., wrong
color, entity absent, opposite direction).
- **uncertain**: Cannot be confirmed or denied due to occlusion, blur,
out-of-frame content, or insufficient information.
### Judgment Principles
1. **Independence**: Judge each fact independently. Do not let the judgment of
one fact influence another.
2. **Visual evidence only**: Base judgments solely on directly observable
content. Do not use logical reasoning or common-sense inference.
3. **Conservative support**: Only mark as support when the visual content clearly
confirms the fact. When in doubt, use uncertain.
```

Video Verification | Reward Prompt *(continued)*

```
### Visual Facts to Verify
```json
{visual_facts}
```
Reply in JSON format:
```json
[
{
"category": "<fact category>",
"fact": "<complete visual fact>",
"analysis": "<specific analysis based on video/image content>",
"result": "<support or conflict or uncertain>"
},
...
]
```
Ensure valid JSON format. All field values must be strings. Do not include
comments in JSON.
```

Figure 11: Full prompt template for video-only claim verification.

Semantic Matching | Reward Prompt

```
You are a semantic matching expert. For the **{category}** category below,
determine whether each "predicted claim" semantically corresponds to a "reference
claim" describing the same visual fact.
### Matching Rules
1. **Semantic equivalence**: Two claims "match" if they refer to the same
specific visual fact, even with different wording or detail level.
2. **Directional constraint**: Each predicted claim matches at most one reference
claim; however, the same reference claim may be matched by at most one predicted
claim.
3. **No match**: If no semantically equivalent reference claim exists, set
gt_index to null and confidence to 0.0.
4. **Reason first**: The reasoning field must be filled before confidence -- use
a concise phrase (≤15 words) extracting subject + action + difference.
5. **Complete output**: The output array must contain exactly {num_pred}
elements, one per predicted claim.
### Confidence Scoring Guide
| Score Range | Meaning | Example |
|---------|------|------|
| 0.9-1.0 | Same fact, different wording | "vehicles moving on the road" ↔ "cars
driving on the street" |
| 0.7-0.8 | Same core fact, detail differences do not affect fact identification |
"a large brown dog" ↔ "a brown Labrador" |
| 0.5-0.6 | Same core entity/action, but missing key distinguishing attributes | "a
cat" ↔ "a black cat with white paws" |
| 0.0-0.4 | Different facts or unrelated | "the sky is blue" ↔ "a dog is running" |
**Important rule**: If the predicted claim omits key distinguishing attributes
(color, quantity, size, material, etc.) from the reference claim, confidence
should NOT exceed 0.6.
--
### Predicted Claims ({num_pred} items)
{pred_claims_json}
### Reference Claims ({num_gt} items)
{gt_claims_json}
### Output Requirement
Output a JSON array containing exactly **{num_pred}** elements, one per predicted
claim:
```json
[
{"pred_index": 0, "gt_index": "<int or null>", "reasoning": "<≤15 word concise
reasoning>", "confidence": "<0.0-1.0>"},
...
]
```
```

Figure 12: Full prompt template for per-category one-to-one semantic matching.

Holistic Scoring | Baseline Prompt

```
### System Message
You are a strict video-caption quality judge.  Evaluate the candidate caption
holistically against both the video and the reference caption, and output exactly
one valid JSON object containing a single overall quality score.  Treat the video
as the primary source of factual evidence and the reference caption as auxiliary
guidance for identifying salient content.
### User Prompt Template
Watch the full video carefully and directly evaluate the overall quality of the
candidate video caption below using the video and the corresponding reference
caption.
### Reference Caption
{reference_caption}
### Candidate Caption
{caption}
### Scoring Criteria
1.  **Factuality**:  Whether the candidate caption faithfully describes the visible
video content.  Consider entities, actions, attributes, spatial relations, event
order, timing, scene changes, and other explicitly stated details.  Unsupported,
contradictory, or temporally incorrect statements should lower the overall score.
2.  **Informativeness**:  Whether the candidate caption covers the important
content shown in the video and described by the reference caption, including
salient subjects, key attributes, actions and events, scene context, interactions,
spatial relations, and temporal changes.
3.  **Overall score**:  Assess the caption as a whole and assign a single quality
score.
### Judgment Principles
- Base the evaluation on the video and the corresponding reference caption.
However, if the reference caption conflicts with clearly visible video content,
follow the video.  Do not reward polished style, fluency, confident wording, or
rhetorical detail by itself.
- Verify explicit timestamps, event order, duration, phases, camera changes, and
scene transitions against the video and the reference caption.  Any inconsistency
should lower factuality.
- Focus on salient content, including the main subjects, key attributes, central
actions and events, important interactions, spatial relations, and temporal
changes.  Exhaustive descriptions of minor background details are not required.
- When the available evidence does not support a statement in the candidate
caption, do not credit it as factually correct.
- Penalize repetition, boilerplate, off-topic content, self-evaluation, and
unsupported filler.  These issues should lower the overall score.
- Before assigning the score, check for material factual errors and salient
omissions.  Base the final score on these issues.
### Score Anchors
- 0.9-1.0:  Factually accurate and covers nearly all important content.
- 0.7-0.8:  Mostly accurate and informative, but misses some important details or
contains mild redundancy.
- 0.5-0.6:  Captures part of the content but misses substantial information or
contains minor factual errors.
- 0.3-0.4:  Contains clear factual errors or provides an overly generic and
incomplete description.
- 0.0-0.2:  Mostly inconsistent with the video and reference caption, or provides
almost no useful description.
### Output Requirement
Output exactly one valid JSON object and no additional text:
```json
{
  "score":  <number from 0.0 to 1.0>,
  "analysis":  "<one-sentence justification for the score>"
}
```
```

Figure 13: Full prompt template for the video- and reference-grounded Holistic reward baseline.

Reference-matching Judge | Baseline Prompt

```
### System Message
You are a strict caption-consistency judge.  You cannot see the video; score only
against the reference caption, use no world knowledge, and output one JSON object.
### User Prompt Template
You are given a reference caption and a candidate caption for the same video.
Treat the reference caption as the complete and only evaluation target, and judge
the candidate solely by comparing it against the reference.
### Reference Caption (the complete evaluation target)
{reference}
### Candidate Caption (to be evaluated)
{caption}
### Scoring Dimensions
1.  **consistency**:  How much of the candidate caption is supported or entailed by
the reference caption.  Content in the candidate that is NOT stated in the
reference should lower this score, even if it might plausibly be true.  You must
NOT use world knowledge or outside assumptions to credit such content.
2.  **coverage**:  How much of the reference caption's content is recovered by the
candidate caption.  Higher when the candidate reflects more of the reference's
subjects, attributes, actions, relations, and events.
### Judgment Principles
- The reference caption is the only source of truth.  Do NOT infer correctness from
world knowledge, plausibility, fluency, or confident wording.
- Candidate details that are absent from the reference receive NO consistency
credit, even if they may be real; treat them as unsupported.
- Do not reward polished style or verbosity; the more unsupported content a
candidate adds, the lower its consistency should be.
- Judge the two dimensions independently:  consistency is "is what the candidate
says supported by the reference", coverage is "does the candidate recover what the
reference says".
### Score Anchors (apply to each dimension)
- 0.9-1.0:  Almost fully supported by / recovers almost all of the reference.
- 0.7-0.8:  Mostly aligned, with some unsupported content or some missed reference
content.
- 0.5-0.6:  Partially aligned; notable unsupported additions or notable omissions.
- 0.3-0.4:  Weak alignment; much content is unsupported or most reference content
is missing.
- 0.0-0.2:  Largely unrelated to the reference.
### Output Requirement
Output only one valid JSON object and no extra text:
```json
{
"consistency":  <number from 0.0 to 1.0>,
"coverage":  <number from 0.0 to 1.0>,
"analysis":  "<one-sentence reason>"
}
```
```

Figure 14: Full prompt template for the Reference-matching reward baseline. The judge receives only the candidate and reference captions.

F.2 Evaluation Prompts

Figure 15 shows the caption-only prompt used for direct caption evaluation and Prism Stage 1. The fixed Stage 2 answerer uses Figure 16 for multiple-choice questions and Figure 17 for open-ended questions. Figure 18 gives the open-ended answer evaluation prompt.

Caption Generation | Evaluation Prompt

```
Provide a detailed and comprehensive description of the video.  The description
should be sufficiently detailed and rich.
```

Figure 15: Caption-only prompt used for direct caption evaluation and Prism Stage 1.

Prism Stage 2 | Multiple-choice QA

```
Given the video caption below, carefully analyze the content and answer the
multiple-choice question.  The prompt provides the generated caption, the benchmark
question, and the answer options.  The QA model must answer directly with the
option letter from the given choices.
```

Figure 16: Stage 2 prompt used by the fixed QA model for multiple-choice questions in Prism-style evaluation.

Prism Stage 2 | Open-ended QA

```
Given the video caption below, carefully analyze the content and answer the
open-ended question.  The prompt provides the generated caption and the benchmark
question.  The QA model must answer directly with a concise free-text response
based only on the caption, including the information needed to answer the question
without a separate reasoning trace.
```

Figure 17: Stage 2 prompt used by the fixed QA model for open-ended questions in Prism-style evaluation.

Prism Scoring | Open-ended Answer Judge

```
### System Message
You are a question-answering evaluation judge.  Evaluate the predicted answer
against the reference answer in the context of the question.  Judge semantic
correctness rather than exact wording, and output one valid JSON object.
### User Prompt
Determine whether the predicted answer correctly answers the question according to
the reference answer below.  Accept paraphrases and semantically equivalent
answers.  Mark the answer incorrect if it contradicts the reference, omits
information necessary to answer the question, or does not answer the question.  Do
not reward fluency or verbosity by itself.
### Question
{question}
### Reference Answer
{reference_answer}
### Predicted Answer
{predicted_answer}
### Output Requirement
Output exactly one valid JSON object and no additional text:
```json
{
  "correct":  <true or false>,
  "reason":  "<one-sentence justification>"
}
```
```

Figure 18: Open-ended answer evaluation prompt for MMVU Open. The judge compares the predicted and reference answers in the context of the question.

G Data Privacy and Content Screening

Public benchmarks. EventHallusion, VCapsBench, DREAM-1K, VideoMME, MMVU, MVBench, MotionBench, TOMATO, and PerceptionTest are publicly released research benchmarks distributed under their respective terms. We use them without modification and do not perform additional screening beyond what their creators applied.

Training videos. All training videos are sourced from publicly available datasets and web platforms that permit research use. We did not collect new video data from human subjects. The videos are used solely for model training and evaluation; we do not attempt to identify, track, or profile individuals appearing in the videos.