DiffCap-RL: Differential QA Rewards for Dense Image and Video Captioning

Kaixun Jiang[1], **Zhihang Liu**[3], **Jiyuan Fu**[2], **Yuzheng Wang**[4 ⋆], **Pandeng Li**[4], **Junjie Zhou**[4], **Chen-Wei Xie**[4], **Yun Zheng**[4], **Wenqiang Zhang**[1,2 †]
[1]College of Intelligent Robotics and Advanced Manufacturing, Fudan University
[2]Shanghai Key Lab of Intelligent Information Processing,
College of Computer Science and Artificial Intelligence, Fudan University
[3]University of Science and Technology of China [4]Alibaba Group

Abstract

Dense image and video captioning demands exhaustive visual details while strictly maintaining factuality, forcing a severe trade-off between descriptive density and hallucination. Supervised fine-tuning (SFT) struggles here: for dense captions, its token-level loss dilutes effective signals while failing to reinforce specific fine-grained dimensions. Reinforcement learning (RL) is a promising alternative but remains bottlenecked by reward design: coarse-grained VLM judges suffer from reward drift, and static QA rewards quickly saturate as the policy improves. To address this, we propose DiffCap-RL, an RL framework combining a stable offline Pre-Scorer with an adaptive online Diff Scorer. We utilize the Pre-Scorer as a stable anchor for broad visual grounding via automated QA pairs. Once it loses discriminative power, Online Diff Scorer dynamically extracts semantic disagreements between policy rollouts. By decomposing these disagreements into targeted conflict questions to suppress hallucinations and extra questions to reward valid new details, the Diff Scorer provides reliable, non-saturating supervision at the policy frontier. We validate our reward quality from three perspectives: human alignment, online RL signal quality, and BoN selection. Across multiple image and video captioning benchmarks, our DiffCap-RL delivers large gains and consistently outperforms state-of-the-art baselines at similar or larger scales. Our page: https://deep-kaixun.github.io/DiffCapRL-page/

1 Introduction

Dense image and video captioning is an important source of supervision for multimodal pre-training, benefiting both Vision-Language Models (VLMs) [10, 1] and Text-to-Image/Video generative models [32, 28, 14, 16]. The core objective of this task is to provide detailed visual narratives as comprehensively as possible (*e.g.*, objects, attributes, relations, background events) while strictly maintaining factual accuracy. Unlike tasks that rely heavily on language priors, this requires meticulous, visually grounded generation where every claim must be supported by the visual input. This high-fidelity, long-form generation poses a significant challenge, forcing existing methods [24, 34] into a trade-off between descriptive density and factuality: pushing for richer descriptions often incurs more severe hallucinations.

Most existing captioners [31, 41, 37] rely on supervised fine-tuning (SFT) using teacher-generated (e.g., GPT [25], Gemini [10]) or human-written data. However, SFT exhibits a clear ceiling for dense captioning. First, it tends to inherit omissions and hallucinations from imperfect training sources. Second, standard token-level cross-entropy distributes the supervision signal evenly across the entire

⋆ Project leader. † Corresponding author.

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

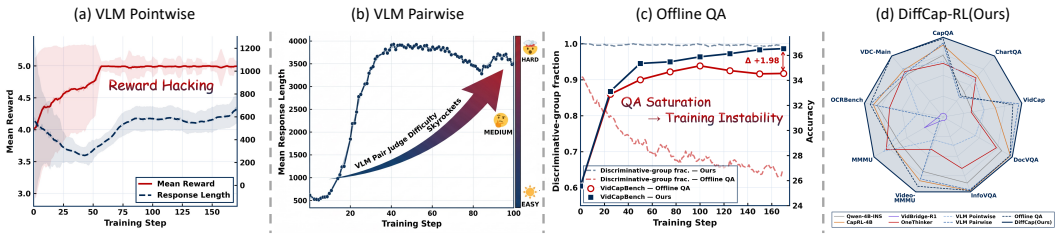

Figure 1: Failure modes of three open-ended caption rewards versus DiffCap-RL. (a) VLM Pointwise: Coarse-grained signals rapidly induce reward hacking. (b) VLM Pairwise: As the output length increases, the evaluation difficulty for the judge escalates, degrading the reward reliability. (c) Offline QA: Relying on predefined QA accuracy as a reward causes the fraction of discriminative groups to gradually decline; leading to performance fluctuations in the later stages of training. (d) DiffCap-RL (Ours): Our approach achieves superior performance across image and video captioning benchmarks.

sequence, overwhelmingly dominated by frequent lexical choices and stylistic tokens. Furthermore, since open-ended dense captions admit many valid alternatives, SFT neither explicitly rewards correct, newly generated details nor directly penalizes visually unsupported claims, ultimately improving stylistic alignment without reliably enhancing fine-grained grounding.

To break the SFT bottleneck, recent work [13, 34, 5, 8, 24, 29] has explored Reinforcement Learning (RL) for dense captioning. Yet, without a unique "ground truth," designing robust rewards remains a critical bottleneck. As shown in Fig. 1, existing mechanisms consistently fail at scale: (1) Pointwise VLM judges [24] provide only coarse-grained signals and quickly lead to reward hacking (Fig. 1a). (2) Pairwise VLM judgments are generally more reliable as they compare relative quality, but their reliability diminishes as captions grow longer and the policy improves: the judge faces increasing difficulty, and the reward signal becomes less reliable, often collapsing into length or positional hacks (Fig. 1b). (3) Offline QA rewards [34, 13, 29] provide fine-grained supervision through pre-defined QA pairs, but also lead to signal saturation. Once the policy saturates the fixed question set, the fraction of discriminative rollout groups drops sharply, causing severe instability in later training (Fig. 1c). Moreover, the fixed QA domains only cover a narrow subset of visual capabilities, leaving broader skills far behind what a more comprehensive reward system could achieve.

To address reward drift and signal saturation, we propose DiffCap-RL, a versatile framework that combines a stable offline Pre-Scorer with an online Diff Scorer. The Pre-Scorer provides a reliable reward anchor via automatically constructed QA pairs for broad visual coverage. As the policy improves, this fixed question set loses discriminative power. The Diff Scorer complements it by dynamically constructing new *conflict* and *extra* questions based exclusively on the semantic disagreements between rollouts at the policy frontier. The key advantage of this differential approach is that evaluating only the difference drastically shortens the context. By filtering out the lengthy, overlapping global content, it prevents the judge from being distracted, making it much easier and more accurate to decompose complex divergences into fine-grained, atomic-level claims. By combining broad offline grounding with adaptive online differential supervision, DiffCap-RL provides a fine-grained, reliable, and non-saturating training signal. We validate DiffCap-RL through human alignment, online RL signal quality, BoN selection, and dimension-wise caption analysis [18]. Models trained with DiffCap-RL achieve leading results on the evaluated captioning benchmarks and deliver substantial gains over their same-size Instruct backbones on multiple downstream VQA tasks. Our contributions are as follows:

- We identify a key bottleneck of RL for dense captioning: existing pointwise, pairwise, and static QA rewards all suffer from *reward drift* or *signal saturation* as training scales.

- We propose **DiffCap-RL**, an online RL framework that combines an offline Pre-Scorer with an online Diff Scorer, decomposing rollout disagreements into *conflict* and *extra* questions for faithful yet detailed caption optimization.

- DiffCap-RL delivers SOTA results on image and video captioning benchmarks, while showing strong human alignment and effectiveness in online RL optimization and test-time scaling selector.

2 Related Work

Dense Image and Video Captioning. Dense captioning is a foundational vision-language task for both images and videos, since caption corpora are widely used to align visual and linguistic spaces during VLM pre-training [1]. Existing caption-specialized models mostly improve quality through data scaling [3, 4, 41, 37]: either collecting costly human annotations, or building *distill-and-refine* pipelines that use strong VLMs. This recipe becomes even more expensive for video, where captions must cover temporal order, camera motion, action transitions, and event structure in addition to objects and scenes. Despite steady progress, SFT-heavy pipelines inherit teacher omissions and hallucinations, and token-level cross-entropy provides weak supervision for the sparse factual details that determine dense-caption quality. DiffCap-RL instead optimizes captions through an offline Pre-Scorer and an online Diff Scorer, providing per-claim feedback; under the same backbone and comparable data budget, it substantially outperforms SFT with fewer training steps.

Reward Design for Open-Ended Captioning. RL has shown strong potential on verifiable tasks [11], but captioning is open-ended without a unique answer while its stated facts remain visually checkable, making reward design central. Direct VLM/LLM judging is the simplest route: pointwise judges, used in video caption RL such as [24, 5, 8], are easily hacked by plausible but weakly grounded captions, while pairwise judges remain vulnerable to length, position, and reference biases [33]. Such holistic signals are too coarse for long captions, where quality depends on sparse factual errors and omissions. Finer-grained rewards instead decompose evaluation into probes: CapRL [34] uses pre-defined visual MCQs, CCCaption [29] adds caption-decomposition correctness checks, and RubiCap [13] aggregates LLM-judged rubric items. These QA/rubric rewards are more reliable, but their supervision is bounded by preset QA or rubric dimensions; once easy probes are solved, rollout rewards become indistinguishable and the RL signal collapses, echoing discrimination collapse in open-ended agent RL [40]. DiffCap-RL addresses this with a stable offline Pre-Scorer plus an online Diff Scorer, generating new visually verifiable questions from rollout disagreements whenever static probes lose resolution.

3 Method

3.1 Preliminaries

Given multimodal inputs x (images or videos), a VLM π_θ produces a caption $c \sim \pi_\theta(\cdot \mid x)$ that should densely and faithfully compress the visual evidence in x into language. For Caption RL, GRPO [11] samples a rollout group $\{c_1, \ldots, c_K\}$ for each x, scores each rollout with reward r_k, and optimises π_θ via group-normalised advantages and a KL regulariser to a frozen reference π_{ref}:

$$\hat{A}_k = \frac{r_k - \text{mean}(\mathbf{r})}{\text{std}(\mathbf{r}) + \epsilon}, \qquad \mathcal{L}(\theta) = -\mathbb{E}\left[\sum_{k=1}^{K} \hat{A}_k \log \pi_\theta(c_k \mid x)\right] + \beta\, D_{\text{KL}}(\pi_\theta \,\|\, \pi_{\text{ref}})\,. \tag{1}$$

Because $\hat{A}_k$ is derived from group-level standardisation rather than absolute values, the overall scale of r is irrelevant; what fundamentally matters is how informatively r captures the fine-grained relative differences among rollouts at the policy's current ability frontier.

Unlike math or code generation, dense captioning lacks a deterministic verifier and admits diverse valid outputs. Moreover, the open-ended nature of visual understanding makes it impossible to exhaustively predefine all "golden rules" or evaluation criteria a priori; any static set of metrics inevitably faces reward saturation as the policy advances. Consequently, designing a robust reward r ideally requires balancing two key properties: (i) **verifiable grounding**—providing objective, multi-dimensional verification of visual facts to mitigate reward drift caused by surface heuristics (e.g., length or position); and (ii) **frontier discrimination**—offering fine-grained signals that adaptively track the policy's evolving capabilities to prevent reward saturation.

Evaluated against these desiderata, current open-ended reward paradigms reveal a fundamental limitation. Dynamic VLM-based judges (whether pointwise or pairwise) often collapse evaluation into an opaque black box. Because their signals are entangled and difficult to verify, they inevitably leak stylistic preferences and position biases, violating verifiable grounding. Conversely, while offline QA evaluations provide a robust, hack-resistant signal, their static scope often leads to **discriminative**

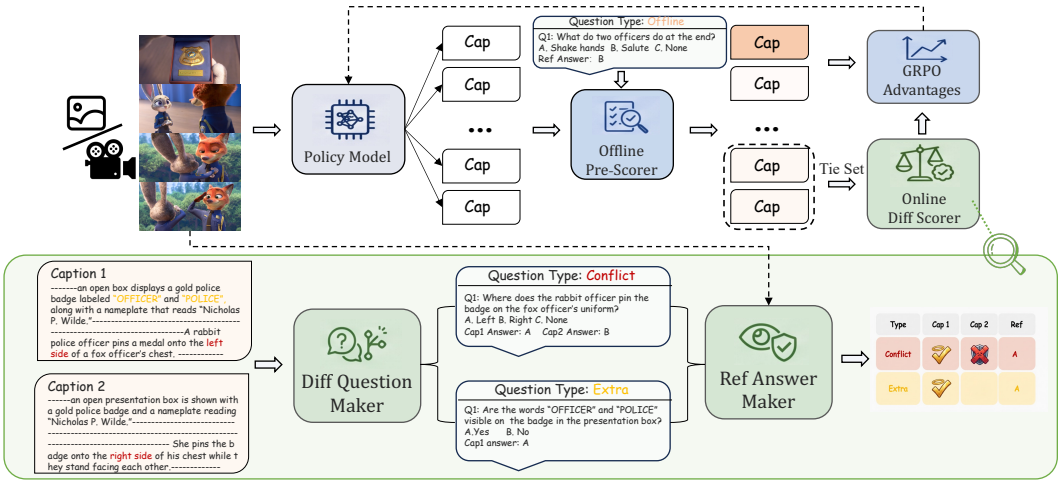

Figure 2: **Overview of the DiffCap-RL framework.** The Offline Pre-Scorer evaluates policy rollouts against a static pool of visually grounded questions, establishing a robust reward baseline. Rollouts with indistinguishable scores are clustered into a tie set. **(Bottom)** The Online Diff Scorer processes these tied rollouts by dynamically identifying factual divergences. It synthesizes targeted questions (*Conflict* and *Extra*) and verifies them against the original media via a Ref Answer Maker, delivering fine-grained, adaptive supervision to overcome discriminative collapse.

collapse: as the policy masters the predefined pool, group-level variance vanishes. In such scenarios, the signal-to-noise ratio plummets, and GRPO's standardisation may inadvertently amplify spurious noise rather than meaningful visual grounding [40], causing optimization to stagnate.

To address this dilemma, we propose DiffCap-RL, which decomposes the reward into two complementary components. First, a stable Offline Pre-Scorer ensures verifiable grounding through a systematically constructed multiple-choice pool (Section 3.2). On top of this, an Online Diff Scorer generates targeted questions based on the rollout group's current disagreements (Section 3.3). Importantly, these disagreements reveal the capability boundaries the policy is exploring at each training step. By delivering precise, verifiable supervision on these evolving frontiers, the online signal helps overcome discriminative collapse. Together, the Offline Pre-Scorer and Online Diff Scorer form a unified reward framework that continuously expands the model's performance limits while keeping generated captions strictly grounded in visual facts. Figure 2 illustrates the full pipeline.

3.2 Offline Pre-Scorer

Motivated by [34], the Pre-Scorer reward r_{off} leverages a static pool of multiple-choice questions (MCQs), $\mathcal{Q}_{\text{off}}(x)$, designed to provide a robust, visually grounded baseline. To automatically construct this pool without human annotation, we execute a streamlined three-stage pipeline: (1) **Multi-dimensional Generation**: Gemini-3-Flash [10] extracts objective, fact-based MCQs from x across diverse visual axes; (2) **Visual-dependency Filtering**: A vision-free LLM removes questions answerable via broad world knowledge without visual evidence; and (3) **Difficulty-Aware Sampling**: The surviving MCQs are sampled based on the current policy's answering accuracy to ensure the final pool $\mathcal{Q}_{\text{off}}(x)$ retains an appropriate difficulty gradient for RL optimization.

During training, r_{off} measures how well a *vision-free* LLM judge $\mathcal{A}$ can answer a set of offline multiple-choice questions using *only* the generated caption c, without access to the image/video x. Here, each q_j denotes a multiple-choice prompt that already includes the question text together with its candidate answer options, and $a_j^\star$ denotes the corresponding correct option. The judge takes the caption c and the MCQ prompt q_j as input, predicts an answer choice $\mathcal{A}(c, q_j)$, and is scored by

$$r_{\text{off}}(c; x) \;=\; \frac{1}{|\mathcal{Q}_{\text{off}}(x)|} \sum_{(q_j, a_j^\star) \in \mathcal{Q}_{\text{off}}(x)} \mathbf{1}\big[\mathcal{A}(c, q_j) = a_j^\star\big]\,. \tag{2}$$

For computational efficiency, all questions associated with the same caption are evaluated jointly in a single batched inference call, yielding an $\mathcal{O}(1)$ number of judge calls per caption.

3.3 Online Diff Scorer

While the static Offline Pre-Scorer provides robust grounding, it inevitably triggers discriminative collapse as the policy improves and rollout scores converge. To restore fine-grained learning signals, the Online Diff Scorer acts as an adaptive supervisor that targets subtle visual disagreements among top-tier rollouts. By dynamically generating questions focused on these unresolved bottlenecks, the Diff Scorer supplies verifiable supervision exactly where the Pre-Scorer loses its resolution.

Stage 1: Tie Clustering. Exhaustive $\mathcal{O}(K^2)$ pairwise comparisons are computationally intractable. To optimize efficiency, we focus the compute budget on the policy's unresolved frontier. Using r_{off} to establish an initial ordering, we group consecutive captions into tied clusters when their unnormalized correct-count gap falls below τ. Within these clusters, we employ a rank-adjacent linear pairing strategy to form the final set $\mathcal{P}(x)$. This design bounds the complexity to $\mathcal{O}(K)$, providing targeted intervention without bottlenecking training throughput.

Stage 2: Diff Question Maker. For each tied pair (c_a, c_b), an LLM synthesizes targeted questions by extracting salient factual disagreements. It formulates these into two types: *Conflict* (mutually exclusive visual claims) and *Extra* (details unique to one rollout). Crucially, the candidate options for these questions are deterministically mapped from the original claims in c_a and c_b. This process inherently establishes a lookup table for each rollout's "answer" to the generated questions, bypassing the need for secondary LLM inference during the later scoring phase. This design eliminates inference noise and ensures that the evaluation is strictly grounded in the rollouts' own assertions.

Stage 3: Ref Answer Maker. To establish ground-truth for reward assignment, we employ a VLM Ref Answer Maker $\mathcal{J}$ that verifies the claims directly against the original media x. Unlike open-ended judges, $\mathcal{J}$ performs atomic visual verification, reducing subjective bias. To mitigate hallucinations, we apply a self-consistency filter: for each question, $\mathcal{J}$ produces M independent answers, and the question is utilized for reward assignment only when all answers reach a strict consensus, yielding the verified ground-truth $a_q^\star$.

Stage 4: Differentiated Diff Scoring. The diff reward prioritizes fine-grained detail while discouraging hallucinated claims. Let $\mathcal{Q}_{\text{diff}}(x)$ denote the robust questions surviving Stage 3. For a rollout c_k, let $a_{k,q}$ denote its inherent claim regarding question q (as mapped in Stage 2).

We design a differentiated scoring schedule based on the cognitive boundary each question probes. *Conflict questions* represent the policy's performance boundaries in resolving visual ambiguities. Thus, matching the ground-truth ($a_{k,q} = a_q^\star$) is rewarded ($+1$), while a mismatch is severely penalized (-1). *Extra questions* represent exploration boundaries. Correctly capturing a verified detail is rewarded ($+0.5$), whereas fabricating an unverified detail is penalized (-0.5) as a hallucination. Omissions receive zero regardless of answer matching. To prevent "length hacking," the scale of extra rewards is constrained. Formally:

$$s(c_k; q) = \begin{cases} +1 & \text{if } q \text{ is conflict and } a_{k,q} = a_q^\star, \\ -1 & \text{if } q \text{ is conflict and } a_{k,q} \neq a_q^\star, \\ +0.5 & \text{if } q \text{ is extra and } a_{k,q} = a_q^\star, \\ -0.5 & \text{if } q \text{ is extra and } a_{k,q} \neq a_q^\star, \\ 0 & \text{if } q \text{ is extra and } c_k \text{ omits the fact.} \end{cases} \tag{3}$$

The online diff reward $r_{\text{diff}}(c_k; x)$ is the average of these scores over $\mathcal{Q}_{\text{diff}}(x)$. The final reward optimized by the RL algorithm is:

$$r_{\text{final}}(c_k; x) = r_{\text{off}}(c_k; x) + \alpha r_{\text{diff}}(c_k; x), \tag{4}$$

where α balances foundational grounding with fine-grained exploration.

4 Experiments

4.1 Experimental Setup

Implementation details. We instantiate DiffCap-RL on Qwen3-VL-4B/8B-Instruct, training for one epoch on $\sim$22k samples ($\sim$7k images, $\sim$15k videos) from open-source training data [8]. Each

sample includes an offline MCQ pool generated by Gemini-3-Flash [10]. During RL, Qwen3-235B [35] acts as the vision-free judge $\mathcal{A}$ and Diff Question Maker, while Qwen3-VL-30B-A3B serves as the Ref Answer Maker using self-consistency ($M = 3$). We optimize using GRPO ($K = 8$ rollouts) and AdamW (learning rate 2×10^{-6}) with a global batch size of 128 and maximum response length of 4096 on 16 A100 GPUs. Unless stated otherwise, $\alpha = 0.3$ and $\tau = 1.0$. More details are provided in Appendix E.

Baselines. We compare our approach against: (i) closed-source VLMs, specifically Gemini-3-Flash [10] and Qwen3.6-Flash [27]; (ii) open-source instruct-tuned models from the Qwen3-VL series [1] (4B, 8B, and 235B); (iii) an SFT baseline (Qwen3-VL-4B) trained on the identical image and video corpus using captions distilled from Qwen3-VL-235B model; and (iv) state-of-the-art (SoTA) captioners, including the CapRL-4B [34] (Qwen3-VL-4B base), OneThinker [8] (Qwen3-VL-8B base), VidBridge-R1 [5] (Qwen2.5-VL-7B base).

Benchmarks and metrics. To fairly demonstrate the model's captioning capabilities (and avoid performance gains from prompt engineering), all prompts are set to *"Describe this image/video in detail."* (i) *Caption benchmarks*: CaptionQA (score) [36] for images, VDC [2] (score, accuracy) and VidCap (Acc. and Prec.) [6] for videos, and CAPability [18] for per-dimension Precision and Hit across 13 fine-grained axes. We use Qwen2.5-VL-72B as a caption-only QA solver for CaptionQA and Llama-3.1-8B-Instruct for caption evaluation on VDC. VidCap and CAPability use Qwen3.5-Plus [30] as the LLM judge. Metric definitions and comparisons across judges are provided in Appendices C and D, respectively. (ii) *Caption-driven QA probes*, inspired by prism [26]: each generated caption is fed to a text-only Qwen3-235B [35] solver that answers standard VQA benchmarks via lmms-eval [39]. This indirectly measures the preservation of visual information across ChartQA [21], DocVQA [23], InfoVQA [22], MMMU [38], OCRBench [17], ScienceQA [19] for images, and Video-MMMU* (Perception track) [12], MVBench [15], VideoMME [9] for videos.

4.2 Main Results

Image captioning and downstream image QA (Tab. 1). DiffCap-RL-4B lifts CaptionQA-Overall to 92.17, achieving a $+6.69$ absolute gain over its Qwen3-VL-4B-Instruct base and a $+2.90$ improvement over the strongest CapRL-4B. Notably, as CaptionQA is a dense-captioning benchmark featuring an average of 12 QA pairs per case, our model's superior results across all sub-domains demonstrate its exceptional ability to capture and provide fine-grained visual details. This performance gain further transfers to text-centric QA probes: DiffCap-RL-4B reaches 88.33 on DocVQA and 77.02 on InfoVQA, both surpassing Gemini-3-Flash. Furthermore, DiffCap-RL-8B sets new performance benchmarks on ChartQA, DocVQA, and InfoVQA, consistently outperforming all compared peer models.

Video captioning and downstream video QA (Tab. 2). DiffCap-RL-4B leads on every video caption-quality metric (*VidCap*, *VDC*) . On *VidCap*, which demands both coverage and per-claim correctness, our model reaches 36.53 in Accuracy (a $+4.52$ gain over the significantly larger Qwen3-VL-235B) and 65.39 in Precision. On *VDC*, a benchmark stressing fine-grained dense descriptions, the gains over Qwen3-VL-235B are most pronounced on Camera-Acc ($+7.93$) and Detail-Acc ($+6.78$), confirming our method's superiority in capturing visual micro-facts. These caption gains transfer to caption-driven VQA: DiffCap-RL-4B sets a new best on Video-MMMU* ($+7.34$ over its 4B base), demonstrating that the generated captions retain the perceptual evidence required by downstream solvers. By contrast, prior RL methods like VidBridge-R1 suffer from severe reward hacking, often collapsing into terse keywords without true visual grounding, resulting in significantly lower performance.

Per-dimension capability. DiffCap-RL establishes a new state-of-the-art across comprehensive evaluations on CAPability [18]. Notably, DiffCap-RL achieves the highest Average Hit rates (67.3% for 4B), significantly outperforming their base models (Qwen3-VL at 52.6% and 53.9%) as well as strong proprietary models like Gemini-3-Flash (61.6%). Beyond general performance, the most substantial improvements manifest in notoriously challenging, long-tail dimensions that require fine-grained perceptual and temporal grounding. For instance, in *spatial relation* (e.g., distinguishing complex geometric layouts like left/right orientations), DiffCap-RL-4B boosts the Hit rate to 83.8%, a massive gain over the base models' $\sim$69%. Similarly, on dynamic visual tasks, DiffCap-RL-8B

Table 1: Image caption quality and caption-driven image QA. CaptionQA is reported by domain and overall score; downstream QA metrics evaluate whether a text-only solver can answer visual questions from the generated caption alone. Best values are in bold; rows shaded blue mark our methods, with Δ rows reporting absolute change against the same-size Qwen3-VL Instruct base.

Method	Param	Training	Caption QA					Image QA					
			Natural	Document	E-comm	Embodied	Overall	ChartQA	DocVQA	InfoVQA	MMMU	OCRBench	ScienceQA
Gemini-3-Flash	–	–	89.01	82.58	92.15	87.33	87.33	48.04	84.27	72.66	**64.00**	**83.80**	**92.97**
Qwen3-VL-235B	235B	Instruct	87.25	89.07	93.47	84.23	87.97	43.08	85.50	74.70	59.11	81.80	91.40
Qwen3.6-Flash	–	–	87.62	88.66	93.69	83.22	87.68	43.76	86.61	74.97	62.56	81.90	92.54
Qwen3-VL-8B-INS	8B	Instruct	85.84	85.10	93.01	82.60	85.97	44.30	83.10	71.30	58.33	80.40	91.20
Qwen3-VL-4B-INS	4B	Instruct	85.54	84.72	92.83	81.66	85.48	46.00	81.90	71.60	58.33	77.50	90.61
SFT (Qwen3-VL-4B)	4B	SFT	87.53	84.83	94.04	83.48	86.76	50.00	78.22	64.54	56.78	75.10	90.49
OneThinker	8B	RL	79.02	83.07	88.76	75.90	80.96	48.50	78.70	62.70	58.78	75.10	91.30
CapRL-4B	4B	RL	89.60	88.52	93.76	86.97	89.27	48.80	87.20	76.32	58.56	78.20	90.75
DiffCap-RL-4B	4B	Ours	**92.02**	**93.12**	**96.81**	**88.65**	**92.17**	52.36	88.33	77.02	59.22	79.90	91.90
Δ (vs Qwen3-VL-4B-INS)			+6.48	+8.40	+3.98	+6.99	+6.69	+6.36	+6.43	+5.42	+0.89	+2.40	+1.29
DiffCap-RL-8B	8B	Ours	90.86	92.36	96.72	87.59	91.35	**52.64**	**89.80**	**78.04**	59.11	81.60	91.86
Δ (vs Qwen3-VL-8B-INS)			+5.02	+7.26	+3.71	+4.99	+5.38	+8.34	+6.70	+6.74	+0.78	+1.20	+0.66

Table 2: Video caption quality and video QA. VidCap reports per-question Accuracy and per-claim Precision; VDC reports Score/Acc. pairs for detailed, main, and camera-oriented dense captioning, and the Δ rows follow the same Score/Acc. order.

Method	Param	Training	VidCap		VDC (Score/Acc.)			Video QA		
			Acc.	Prec.	Detail	Main	Camera	Video-MMMU*	MVBench	VideoMME
Gemini-3-Flash	–	–	26.05	56.30	2.65/53.51	2.81/56.71	2.37/47.69	71.67	**60.55**	**64.52**
Qwen3-VL-235B	235B	Instruct	32.01	61.84	2.89/58.19	3.04/61.21	2.48/49.99	73.33	56.18	62.38
Qwen3.6-Flash	–	–	24.74	56.31	2.76/55.80	2.88/57.96	2.43/48.76	69.67	56.45	60.33
Qwen3-VL-8B-INS	8B	Instruct	27.42	58.59	2.87/58.03	3.07/61.80	2.55/51.19	69.33	55.43	61.00
Qwen3-VL-4B-INS	4B	Instruct	25.56	57.18	2.83/56.95	3.03/60.82	2.54/51.14	68.33	50.50	60.18
SFT (Qwen3-VL-4B)	4B	SFT	31.50	60.25	2.90/58.28	3.02/60.64	2.47/49.40	70.00	50.98	60.26
OneThinker	8B	RL	25.22	59.75	2.79/56.43	2.82/56.73	2.55/51.22	60.33	54.43	57.48
VidBridge-R1	7B	RL	19.80	49.20	2.38/48.53	2.47/49.99	2.39/48.46	50.00	40.38	46.22
DiffCap-RL-4B	4B	Ours	**36.53**	**65.39**	**3.24/64.97**	**3.29/65.80**	**2.91/57.92**	**75.67**	58.03	63.70
Δ (vs Qwen3-VL-4B-INS)			+10.97	+8.21	+0.41/+8.02	+0.26/+4.98	+0.37/+6.78	+7.34	+7.53	+3.52
DiffCap-RL-8B	8B	Ours	35.06	65.19	3.13/62.86	3.16/63.54	2.84/56.84	74.00	58.52	64.03
Δ (vs Qwen3-VL-8B-INS)			+7.64	+6.60	+0.26/+4.83	+0.09/+1.74	+0.29/+5.65	+4.67	+3.09	+3.03

achieves remarkable breakthroughs: it reaches a Hit rate of 37.3% on *dynamic object number* and 59.7% on *camera movement*, more than tripling and nearly doubling the performance of prior video captioners like VidBridge-R1 (6.7% and 30.9%).

SFT vs. RL: fluency without grounding (Tab. 1). SFT lifts Caption-Overall by $+1.28$ but *regresses* on DocVQA (-3.68), InfoVQA (-7.06), and OCRBench (-2.40); DiffCap-RL-4B instead improves over the same Instruct base ($+6.43$, $+5.42$, $+2.40$). The contrast reflects the impact of supervision granularity: SFT imitates teacher captions holistically, inadvertently inheriting their omissions and hallucinations. In contrast, DiffCap-RL provides per-claim QA verification and dimension-level rewards via the Diff Scorer's Ref Answer Maker. By explicitly supervising every visual axis and individually rewarding each correct atomic detail, our approach ensures that caption quality and downstream QA performance improve synergistically rather than trading off.

Backbone scaling: 4B vs. 8B. We find caption quality grows modestly with backbone size on the Qwen3-VL family. A plausible reading is that scaling chiefly enlarges the LLM component, whereas captioning is a VLM-specific perception skill whose ceiling is more tightly bounded by the availability of high-quality caption–image/video pairs, which remain non-trivial to curate at comparable scale. Since RL elicits rather than injects visual capability, the same pattern carries over to DiffCap-RL: the 4B variant slightly leads on caption quality, the 8B on broader QA probes, and both consistently surpass their Instruct bases (Δ rows).

4.3 What Makes a Good Caption Reward?

We next examine why DiffCap-RL behaves differently from prior caption rewards. Beyond final benchmark scores, we evaluate whether a reward can serve three roles: (i) align with human preferences (Tab. 3), (ii) provide a non-degenerate RL training signal (Tab. 4, Fig. 4), and (iii) act as an inference-time selector (Fig. 5). We verify each in turn, briefly on (i) and (iii) and in depth on (ii).

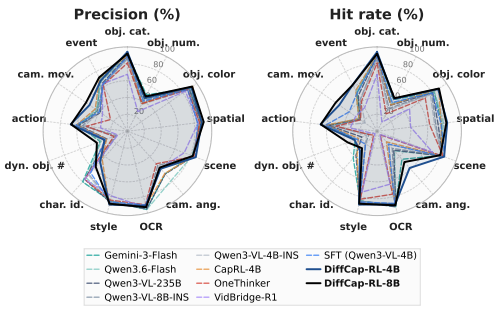

Figure 3: Per-dimension capability profile on Capability. Left: Precision (per-claim factuality). Right: Hit (per-fact recall). Each radar reports 13 fine-grained visual axes across 11 captioners.

Table 3: Human-alignment evaluation. CapArena is a human-annotated image-caption arena; caption level reports average agreement with human pairwise decisions, and model level reports Spearman's ρ between reward-induced model ranking and human ranking. VideoCapArena is our human-annotated video-caption arena. All VLM-based rewards use Gemini-3-Flash as the judge.

Reward setting	CapArena		VideoCapArena
	Caption level	Model level	Caption level
VLM pointwise judge	40.50	0.8590	10.00
VLM pairwise judge	64.80	0.9380	55.00
Offline QA	58.20	0.9428	49.00
DiffCap-RL	**65.40**	**0.9516**	**64.00**

Table 4: Reward-design ablation. The full DiffCap-RL reward combines an Offline Pre-Scorer with an Online Diff Scorer, attaining the best scores on every axis (caption quality, image and video QA probes, video caption quality).

Method	Caption QA					Image QA		Video QA	VidCap	
	Natural	Document	E-comm	Embodied	Overall	ChartQA	OCRBench	Video-MMMU*	Acc.	Prec.
Qwen3-VL-4B-INS (base)	85.54	84.72	92.83	81.66	85.48	46.00	77.50	68.33	25.56	57.18
VLM pointwise judge	85.43	80.04	87.69	81.31	83.13	48.16	69.60	57.00	28.10	63.85
VLM pairwise judge	90.57	89.83	96.40	85.88	90.02	45.04	78.80	71.67	35.26	62.75
Offline QA	91.17	92.66	96.54	88.17	91.65	45.32	78.80	73.67	34.55	61.77
Diff-QA only	90.84	91.64	96.00	88.25	91.22	50.92	78.00	75.33	34.97	64.46
DiffCap-RL	**92.02**	**93.12**	**96.81**	**88.65**	**92.17**	**52.36**	**79.90**	**75.67**	**36.53**	**65.39**

We compare DiffCap-RL against three reward designs that span the three roles. A *pointwise VLM judge* scores each caption independently under a fixed rubric. A *pairwise VLM judge* compares each sampled caption against a dynamic reference, namely the temperature-0 caption from the current policy, and maps win/tie/loss to $+1/0/-1$; this gives a practical preference baseline without all-pairs comparison or a fixed external reference. *Offline QA* follows CapRL-style static QA supervision.

Reward with Human alignment. We first test whether each reward agrees with humans. CapArena [7] is a human-annotated image-caption arena with pairwise comparisons among multiple models, we further build VideoCapArena with 100 carefully annotated video-caption pairs, covering both human error-correction pairs derived from the same caption and pairs generated by different caption models. For fairness, all VLM-based rewards use Gemini-3-Flash as the judge. DiffCap-RL achieves the strongest alignment on CapArena at both caption level and model level (**65.40/0.9516**), while the pairwise VLM judge is close because short image captions are relatively easy to compare. On the harder VideoCapArena, the gap widens: DiffCap-RL reaches **64.00**/100, $+9.00$ over the pairwise VLM judge, indicating that VLM-based pairwise comparison becomes less reliable as caption length and temporal detail increase.

Reward as RL signal. The ablations suggest that an effective caption reward must provide not only high scores but also *granular* and *stable* training signals. The *pointwise VLM judge* performs worst: since the base model already produces broadly plausible captions, major hallucinations are rare, and the remaining errors are mostly omissions of fine-grained details. A single coarse score is thus easy to exploit, as it rewards general plausibility without encouraging verified coverage. The *pairwise VLM judge* is more effective in the early stage. When sampled captions differ substantially in length and detail, the judge can usually identify the better one, leading to rapid initial gains in Fig. 4. As captions become longer and more similar in quality, however, holistic comparison becomes less reliable and still offers no per-claim credit assignment,

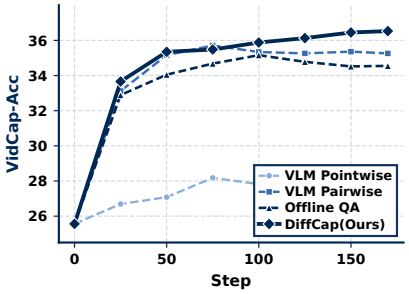

Figure 4: Per-step VidCap-Acc for three baseline rewards (pointwise / pairwise VLM judge, Offline QA) and DiffCap-RL. DiffCap-RL is the only reward that monotonically improves.

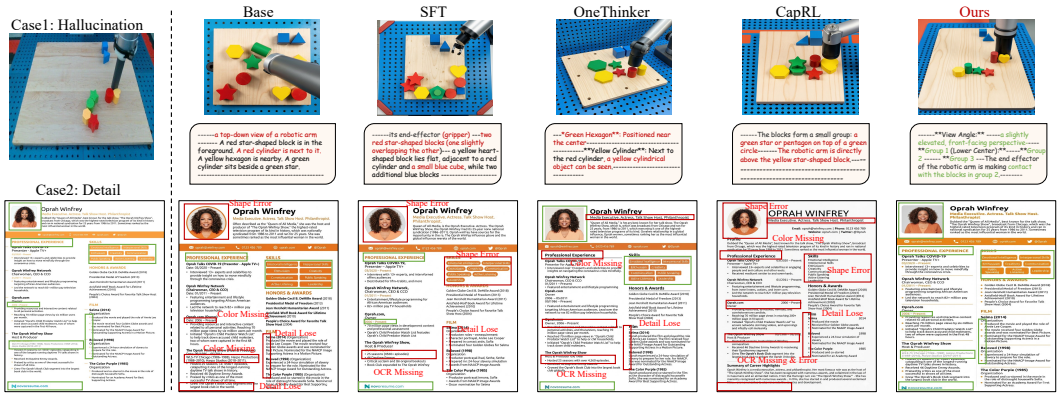

Figure 6: Visual evaluation of caption quality via text-to-image reconstruction. Images are generated by feeding the respective models' output captions into Nano Banana 2 [10]. The visual comparisons demonstrate that our method produces fewer hallucinations in complex scenes (Case 1) and preserves richer details in information-dense images (Case 2). Red boxes highlight errors or omissions, while green boxes indicate accurately captured details.

causing learning to plateau. The *Offline QA* reward adds grounded supervision, but its fixed question pool cannot adapt to the policy's evolving omissions and errors. As a result, the number of effective groups drops sharply during training (Fig. 1c), and performance saturates after step 75. To isolate the effect of the online differential signal, we further evaluate a fourth variant, *diff-QA only*, which keeps the Diff Scorer but removes the Offline Pre-Scorer. Since this setting lacks a native single-caption grounding score, we construct the contrast efficiently by sampling non-repeating caption pairs within each rollout group, rather than enumerating all $\binom{K}{2}$ pairs. This variant targets current caption differences and improves long-tail benchmarks such as ChartQA (50.92) and Video-MMMU* (75.33), but weakens stable coverage of basic skills, as shown by the lower OCRBench score (78.00 vs. 78.80 for Offline QA). DiffCap-RL combines these two QA signals: the Offline Pre-Scorer provides a stable grounding baseline, while the Online Diff Scorer contributes dynamic, per-claim supervision. This combination achieves the best result on every axis in Tab. 4 and is the only method with a monotonically improving VidCap-Acc curve in Fig. 4.

Reward as BoN Selector. A robust caption reward should remain discriminative even when faced with multiple highly similar candidate captions, selecting the one with clearer and more complete visual evidence rather than merely scoring all fluent outputs similarly. We therefore evaluate DiffCap-RL as a selector over $N \in \{1, 2, 4, 8\}$ candidates sampled from a fixed policy. As N increases, the reward consistently selects better captions, monotonically lifting both image and video caption quality (Fig. 5).

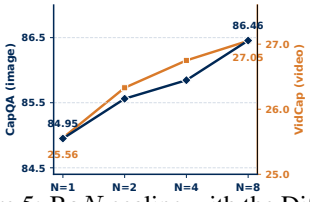

Figure 5: BoN scaling with the DiffCap-RL reward as a caption selector.

4.4 Qualitative Analysis

We present two case studies as shown in Fig 6 by generating reverse-mapped images from captions produced by different models using Nano Banana 2. By comparing the generated images and their correspondence to the source images, we analyze hallucination and descriptive detail. First, in scenes involving complex spatial relations, object counts, and other compositional attributes, our method produces fewer hallucinations than competing approaches. Second, for images with high information density, our method generates captions that capture richer and more fine-grained details. Moreover, we provide more image and video captioning cases and additional cases for diff QA in Appendix F.

5 Conclusion

We presented DiffCap-RL, a novel reinforcement learning framework for dense image and video captioning that resolves the core reward-design tension between grounded verification and frontier-

level discrimination. By coupling a stable offline Pre-Scorer with an adaptive online Diff Scorer, our approach dynamically converts rollout disagreements into targeted conflict and extra questions. This ensures continuous, fine-grained supervision even after static rewards saturate, strictly anchoring the generation to visual evidence. Through extensive evaluations—spanning captioning benchmarks, caption-driven downstream VQA tasks, human alignment, and Best-of-N selection—we demonstrate that differential QA rewards effectively break the trade-off between factuality and descriptive density.

Broader impacts. Better dense captions may improve accessibility, media understanding, and supervision for multimodal models. Potential risks include misuse in surveillance or misleading media pipelines if deployed without privacy, consent, and bias safeguards; downstream deployments should therefore include appropriate data-governance and safety checks.

Acknowledgements. This work was supported by the National Natural Science Foundation of China (No. 62576109), the Scientific and Technological Innovation Action Plan of the Shanghai Science and Technology Committee (No. 25511104402), and the AI for Science Program of the Shanghai Municipal Commission of Economy and Informatization (No. 2025-GZL-RGZN-BTBX-02006).

References

[1] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[2] Wenhao Chai, Enxin Song, Yilun Du, Chenlin Meng, Vashisht Madhavan, Omer Bar-Tal, Jenq-Neng Hwang, Saining Xie, and Christopher D Manning. Auroracap: Efficient, performant video detailed captioning and a new benchmark. *arXiv preprint arXiv:2410.03051*, 2024.

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

[4] Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, et al. Sharegpt4video: Improving video understanding and generation with better captions. *Advances in Neural Information Processing Systems*, 37:19472–19495, 2024.

[5] Xinlong Chen, Yuanxing Zhang, Yushuo Guan, Weihong Lin, Zekun Wang, Bohan Zeng, Yang Shi, Sihan Yang, Qiang Liu, Pengfei Wan, et al. Vidbridge-r1: Bridging qa and captioning for rl-based video understanding models with intermediate proxy tasks. *arXiv preprint arXiv:2506.09079*, 2025.

[6] Xinlong Chen, Yuanxing Zhang, Chongling Rao, Yushuo Guan, Jiaheng Liu, Fuzheng Zhang, Chengru Song, Qiang Liu, Di Zhang, and Tieniu Tan. Vidcapbench: A comprehensive benchmark of video captioning for controllable text-to-video generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 8543–8563, 2025.

[7] Kanzhi Cheng, Wenpo Song, Jiaxin Fan, Zheng Ma, Qiushi Sun, Fangzhi Xu, Chenyang Yan, Nuo Chen, Jianbing Zhang, and Jiajun Chen. Caparena: Benchmarking and analyzing detailed image captioning in the llm era. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14077–14094, 2025.

[8] Kaituo Feng, Manyuan Zhang, Hongyu Li, Kaixuan Fan, Shuang Chen, Yilei Jiang, Dian Zheng, Peiwen Sun, Yiyuan Zhang, Haoze Sun, et al. Onethinker: All-in-one reasoning model for image and video. *arXiv preprint arXiv:2512.03043*, 2025.

[9] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. 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*, pages 24108–24118, 2025.

[10] Google. gemini3. `https://blog.google/products-and-platforms/products/gemini/gemini-3/`, 2025.

[11] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[12] Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*, 2025.

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

[14] Kaixun Jiang, Yuzheng Wang, Junjie Zhou, Pandeng Li, Zhihang Liu, Chen-Wei Xie, Zhaoyu Chen, Yun Zheng, and Wenqiang Zhang. Genagent: Scaling text-to-image generation via agentic multimodal reasoning. *arXiv preprint arXiv:2601.18543*, 2026.

[15] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024.

[16] Zhaohe Liao, Kaixun Jiang, Zhihang Liu, Yujie Wei, Junqiu Yu, Quanhao Li, Hong-Tao Yu, Pandeng Li, Yuzheng Wang, Zhen Xing, et al. Aibench: Evaluating visual-logical consistency in academic illustration generation. *arXiv preprint arXiv:2603.28068*, 2026.

[17] Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024.

[18] Zhihang Liu, Chen-Wei Xie, Bin Wen, Feiwu Yu, Jixuan Chen, Pandeng Li, Boqiang Zhang, Nianzu Yang, Yinglu Li, Zuan Gao, et al. Capability: A comprehensive visual caption benchmark for evaluating both correctness and thoroughness. *arXiv preprint arXiv:2502.14914*, 2025.

[19] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[20] Ziyang Ma, Ruiyang Xu, Zhenghao Xing, Yunfei Chu, Yuxuan Wang, Jinzheng He, Jin Xu, Pheng-Ann Heng, Kai Yu, Junyang Lin, et al. Omni-captioner: Data pipeline, models, and benchmark for omni detailed perception. *arXiv preprint arXiv:2510.12720*, 2025.

[21] Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pages 2263–2279, 2022.

[22] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and CV Jawahar. Infographicvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1697–1706, 2022.

[23] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209, 2021.

[24] Desen Meng, Rui Huang, Zhilin Dai, Xinhao Li, Yifan Xu, Jun Zhang, Zhenpeng Huang, Meng Zhang, Lingshu Zhang, Yi Liu, et al. Videocap-r1: Enhancing mllms for video captioning via structured thinking. *arXiv preprint arXiv:2506.01725*, 2025.

[25] OpenAI. Gpt-5. `https://openai.com/index/introducing-gpt-5/`, 2025.

[26] Yuxuan Qiao, Haodong Duan, Xinyu Fang, Junming Yang, Lin Chen, Songyang Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. Prism: A framework for decoupling and assessing the capabilities of vlms. *Advances in Neural Information Processing Systems*, 37:111863–111898, 2024.

[27] Qwen Team. Qwen3.6-35B-A3B: Agentic coding power, now open to all, April 2026.

[28] Team Seedance, De Chen, Liyang Chen, Xin Chen, Ying Chen, Zhuo Chen, Zhuowei Chen, Feng Cheng, Tianheng Cheng, Yufeng Cheng, et al. Seedance 2.0: Advancing video generation for world complexity. *arXiv preprint arXiv:2604.14148*, 2026.

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

[30] Qwen Team. Qwen3.5: Accelerating productivity with native multimodal agents, February 2026.

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

[32] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

[33] Peiran Wu, Yunze Liu, Zhengdong Zhu, Enmin Zhou, and Junxiao Shen. Ugc-videocaptioner: An omni ugc video detail caption model and new benchmarks. *arXiv preprint arXiv:2507.11336*, 2025.

[34] 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. *arXiv preprint arXiv:2509.22647*, 2025.

[35] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[36] Shijia Yang, Yunong Liu, Bohan Zhai, Ximeng Sun, Zicheng Liu, Emad Barsoum, Manling Li, and Chenfeng Xu. Captionqa: Is your caption as useful as the image itself? *arXiv preprint arXiv:2511.21025*, 2025.

[37] 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. *arXiv preprint arXiv:2501.07888*, 2025.

[38] Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9556–9567, 2024.

[39] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Reality check on the evaluation of large multimodal models, 2024.

[40] Qiang Zhang, Boli Chen, Fanrui Zhang, Ruixue Ding, Shihang Wang, Qiuchen Wang, Yinfeng Huang, Haonan Zhang, Rongxiang Zhu, Pengyong Wang, et al. Arenarl: Scaling rl for open-ended agents via tournament-based relative ranking. *arXiv preprint arXiv:2601.06487*, 2026.

[41] 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 arXiv:2410.02713*, 2024.

Table 5: Training-data ablation. Joint image+video training dominates video-only on every reported caption and QA axis, confirming that the unified pipeline benefits from cross-modality transfer rather than diluting the video signal. The row shaded blue is our default configuration.

Training data	Caption QA					Image QA		Video QA	VidCap	
	Natural	Document	E-comm	Embodied	Overall	ChartQA	OCRBench	Video-MMMU*	Acc.	Prec.
None (Qwen3-VL-4B-INS)	85.54	84.72	92.83	81.66	85.48	46.00	77.50	68.33	25.56	57.18
Video only	91.67	91.59	96.56	88.33	91.53	45.72	78.00	74.33	35.26	64.26
Image & video	**92.02**	**93.12**	**96.81**	**88.65**	**92.17**	**52.36**	**79.90**	**75.67**	**36.53**	**65.39**

Table 6: Reward-weight ablation on α in $r_{\text{final}} = r_{\text{off}} + \alpha\, r_{\text{diff}}$.

α	Caption QA					Image QA		Video QA	VidCap	
	Natural	Document	E-comm	Embodied	Overall	ChartQA	OCRBench	Video-MMMU*	Acc.	Prec.
0.1	91.04	92.47	96.54	87.93	91.49	49.48	**81.30**	73.00	35.30	63.22
0.3	**92.02**	**93.12**	96.81	**88.65**	**92.17**	**52.36**	79.90	**75.67**	**36.53**	**65.39**
0.5	91.22	91.58	**96.85**	88.41	91.50	50.20	80.80	74.67	36.08	64.75

A Overview

This appendix collects details that are useful for reproducing or inspecting DiffCap-RL but are not necessary for the main narrative.

Appendix B provides additional ablations on training data and the Diff Scorer weight.

Appendix C defines the caption evaluation metrics, and Appendix D reports robustness to different judge models on CaptionQA.

Appendix E describes the implementation protocol, reward construction details, prompt templates, and VideoCapArena annotation setup.

Appendix F presents complete qualitative cases for image and video captioning.

Appendix G summarizes limitations and future work.

B Ablation Studies

Joint image+video data (Tab. 5). We use this ablation to test whether DiffCap-RL benefits from a unified image–video captioning distribution. Training only on the video set already improves substantially over the base model, showing that the reward can learn from video supervision alone. Adding the image partition further improves every reported caption and QA axis, including CaptionQA-Overall (91.53 $\rightarrow$ 92.17), ChartQA (45.72 $\rightarrow$ 52.36), Video-MMMU* (74.33 $\rightarrow$ 75.67), and VidCap-Acc (35.26 $\rightarrow$ 36.53). Thus, image data does not dilute the video signal; instead, it transfers fine-grained visual grounding cues such as object detail, spatial layout, OCR, and scene structure to the shared captioning policy.

Diff Scorer weight α (Tab. 6). $\alpha = 0.3$ is the best overall trade-off: $\alpha = 0.1$ under-amplifies the Diff Scorer and inherits the Pre-Scorer plateau (CaptionQA-Overall 91.49, VidCap-Acc 35.30), while $\alpha = 0.5$ over-amplifies the online signal and slightly hurts CaptionQA-Overall (91.50). Although smaller weights preserve OCRBench slightly better, $\alpha = 0.3$ jointly maximizes CaptionQA-Overall (92.17), ChartQA (52.36), Video-MMMU* (75.67), and VidCap-Acc (36.53), confirming a balance between Pre-Scorer stability and Diff Scorer discrimination.

C Evaluation Metrics

CaptionQA. CaptionQA [36] measures caption utility through image-related multiple-choice questions across four domains. We use Qwen2.5-VL-72B-Instruct to answer from captions alone, apply the benchmark's QA scoring rule, and report Overall as the domain scores weighted by their image counts.

Table 7: CaptionQA performance under different judge models.

	QA judge		
Method	Qwen2.5-VL-72B	Qwen2.5-72B	GPT-5.5
CapRL-4B	89.27	88.39	92.83
DiffCap-RL-4B	**92.17**	**90.74**	**94.49**

VDC. VDC [2] evaluates detailed video captions using reference QA pairs. We use Llama-3.1-8B-Instruct for caption-based QA evaluation and report Score (0–5) and Accuracy for the detailed, main-object, and camera subsets.

VidCapBench. VidCapBench [6] evaluates how well video captions convey information covered by reference QA pairs. We report Accuracy (fully correct answers among all questions) and Precision (correct or partially correct answers among non-neutral answers).

CAPability. CAPability [18] assesses fine-grained captioning across 13 evaluated dimensions. We report Precision (correctness of mentioned information) and Hit (the fraction of annotated information correctly conveyed).

D Robustness to Different Judge Models

To assess sensitivity to the evaluation model, we compare DiffCap-RL-4B with CapRL-4B on CaptionQA using Qwen2.5-VL-72B-Instruct, Qwen2.5-72B-Instruct, and GPT-5.5. The two Qwen solvers evaluate the same generated captions, and GPT-5.5 provides an additional cross-family evaluation. All judges answer questions from captions without access to images. As shown in Table 7, DiffCap-RL-4B outperforms CapRL-4B under all three judges, supporting the robustness of its advantage across evaluation model families.

E Additional Implementation Details

Offline QA Maker and Solver prompts. We provide the prompt templates of the Offline QA Maker and the Offline QA Solver in Fig. 14 and Fig. 15, respectively.

Diff Question Maker prompt. We provide the prompt template of the Diff Question Maker in Fig. 16.

VLM Pointwise and Pairwise Judge. To ensure fairness, we use the same model, Qwen3-VL-30b-A3B, for both the VLM pointwise and pairwise judge (the same as our ref answer maker) during online RL training. The prompts used for the pointwise and pairwise judges are shown in Fig. 13 and Fig. 12, respectively.

Ref Answer Maker prompt. We provide the prompt template of the Ref Answer Maker in Fig. 11.

VideoCapArena details. VideoCapArena contains 100 human-annotated pairwise video-caption comparisons designed to stress long, temporally detailed captions. The pairs cover two complementary regimes: error-correction pairs derived from the same caption, where annotators choose between an original caption and a minimally edited version with factual corrections, and model-comparison pairs, where annotators compare captions produced by different captioning systems for the same video. Annotators are instructed to prefer captions that are both faithful and sufficiently detailed, penalizing unsupported events, wrong temporal order, incorrect object attributes, and missing salient actions. This setup makes VideoCapArena harder than short image-caption comparisons and better aligned with the failure modes targeted by the Diff Scorer.

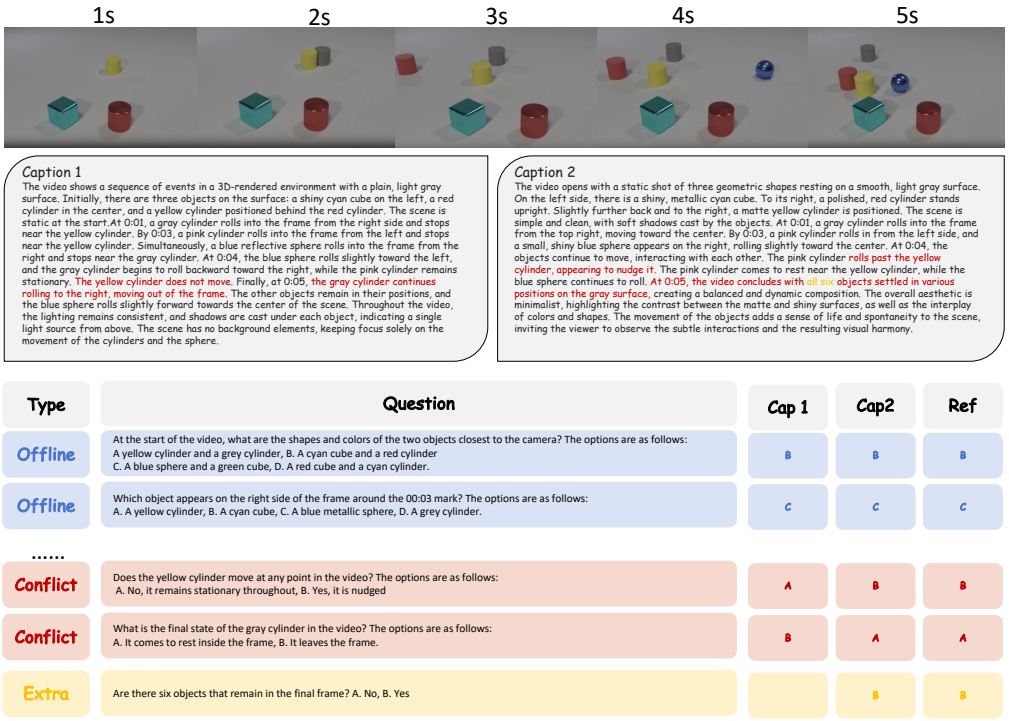

Figure 7: Complete case for DiffCap-RL QA.

F More Case Studies

Fig. 9, Fig. 10, Fig. 7, and Fig. 8 provide full qualitative examples for image and video captioning. They complement Fig. 6 by showing the complete captions and QA-style checks behind the summarized case study.

G Limitations and Future Work

This work focuses on dense captioning for images and videos, where visual details, temporal events, and their relations can be verified through targeted QA. The Diff Scorer cannot directly detect hallucinations or omissions shared by all rollouts, and offline supervision remains limited by QA coverage. A natural next step is audio-visual captioning [33, 20], which is increasingly important for omni-modal assistants and media understanding: many events are defined not only by what is visible, but also by speech, sound sources, acoustic context, and their temporal alignment with the visual stream. Such settings introduce richer cross-modal evidence and make faithful captioning even more challenging and valuable. Importantly, DiffCap-RL is not tied to a specific input modality. Its reward design only requires that disputed caption facts can be converted into verifiable questions and answered against the original media. We therefore expect the framework to extend naturally to omni-modal models by replacing the base captioner and verifier with audio-video-capable models, and by augmenting the QA construction with audio-visual question types. We leave this broader omni-modal extension, together with multilingual and domain-specific evaluations, to future work.



Caption 1

The video begins with a scene set in a cozy kitchen. A person's hands are shown carefully assembling a sandwich on a black frying pan placed on the stove. The background reveals a variety of kitchen items, including a red kettle, several glass jars, and a black plate with a spoon resting on it. These elements suggest a typical home kitchen environment, adding to the homely and inviting atmosphere of the scene. The process of making the sandwich starts with a slice of bread placed in the pan. The person then begins layering the sandwich, starting with slices of cheese. Following the cheese, they add slices of pickles, which are neatly arranged on top. Next, they place slices of jalapeño peppers, adding a touch of spice to the sandwich. The person ensures that the peppers are evenly distributed across the sandwich. Finally, they add slices of fresh tomato on top, completing the layering process. As the video progresses, the person's hands are seen carefully assembling the sandwich, demonstrating precision and attention to detail. The camera provides close-up shots of the sandwich as it takes shape, allowing viewers to see the neat arrangement of ingredients. The final shot showcases the completed sandwich, with all the layers clearly visible, ready to be cooked. Overall, the video captures a simple yet satisfying kitchen activity — making a sandwich — with a focus on the careful layering of ingredients and the inviting atmosphere of a home kitchen. The video concludes with the sandwich fully assembled, ready to be enjoyed.

Caption 2

The video opens with a close-up shot of a black frying pan positioned on a stovetop in a kitchen. A slice of whole wheat bread, with cheese already placed on top, is being prepared into a sandwich. The setting is a homely kitchen environment, with various items like a red kettle, glass jars, and a black plate with a spoon visible in the background, adding to the cozy, everyday atmosphere. A person's hands enter the frame, and they begin assembling the sandwich. First, they carefully place two slices of pickle on top of the cheese, arranging them neatly. Following this, they add a layer of sliced jalapeño peppers twice in succession, distributing them evenly over the pickles each time. Then, the person places two slices of ripe, red tomato on top of the jalapeños. The sandwich ingredients are layered neatly, with the colorful vegetables creating a visually appealing contrast against the cheese and bread. As the video progresses, the camera zooms in slightly on the sandwich, highlighting the vibrant colors and textures of the ingredients, from the green of the pickles and jalapeños to the fresh red of the tomatoes. The final shot shows the fully assembled sandwich in the pan. The kitchen background remains consistent, maintaining the warm and inviting feel throughout the video. Overall, the video captures the simple, satisfying process of making a sandwich, emphasizing the fresh ingredients and the care taken in assembling it. It conveys a sense of home-cooked comfort and the pleasure of preparing a delicious meal.

Type	Question	Cap 1	Cap2	Ref
Offline	What is the color of the kettle located on the left side of the stove? The options are as follows: A. Silver, B. Blue, C. Red, D. Green.	C	C	C
Offline	What is the very last ingredient placed on the sandwich during the video? The options are as follows: A. A slice of sharp cheddar cheese, B. Two slices of tomato, C. Several jalapeño slices, D. Three pickle slices.	B	B	B
.....				
Conflict	When the bread is placed in the frying pan, is cheese already on it? The options are as follows: A. No, cheese is added after the bread is placed in the pan, B. Yes, cheese is already on the bread, C. Neither.	A	B	B
Conflict	How many times is the process of adding jalapeño peppers shown in the video? The options are as follows: A. Once, B. Twice.	A	B	B
Extra	How many pickle slices were added? A. One slice B. Two slices C. Others		B	B
Extra	How many tomato slices were added? A. One slice B. Two slices C. Others		B	B

Figure 8: Complete case for DiffCap-RL QA. Extra questions are reformulated for clearer visualization.

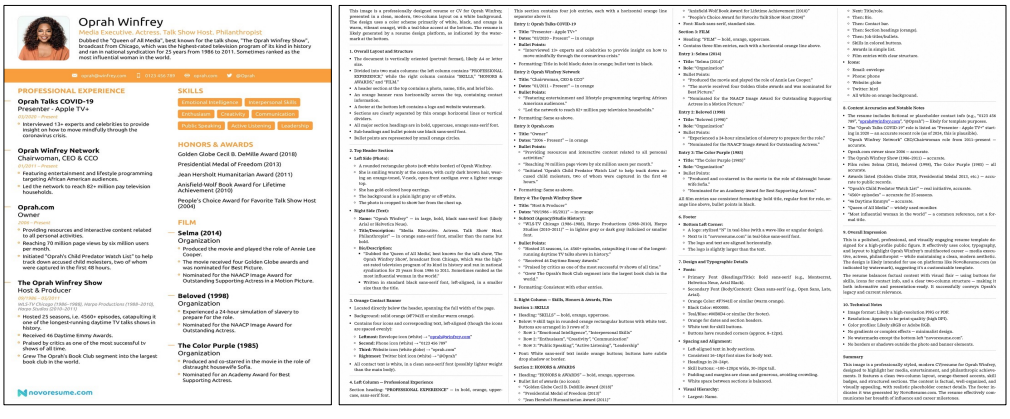

Figure 9: Complete image-captioning example from DiffCap-RL-4B. The image is on the left and the caption is on the right.



Figure 10: Complete video-captioning example from DiffCap-RL-4B. The video is above and the caption is below.

Ref Answer Maker (Video)

```
r'''
# Role
You are an expert in Video Understanding and Evaluation.


# Task
You will be provided with a video and a list of multiple-choice questions.
Your goal is to answer **ALL** questions correctly based on visual evidence.


# Input Questions
{questions_json}


# Instructions (Strictly Follow This Order)


## Part 1: Step-by-Step Reasoning
For **EACH** question ID, you must:
1. **Locate**: Identify the specific time segments in the video relevant to the question.
2. **Analyze**: Describe what you see in those segments.
3. **Eliminate**: Briefly explain why incorrect options are wrong.
4. **Conclude**: State the correct option.


## Part 2: Final JSON Output
[Your Step-by-Step Reasoning here...]


```json
{{
"answers": [
{{"id": 1, "reasoning": "The correct option is A because...", "answer": "A"}},
{{"id": 2, "reasoning": "The correct option is B because...", "answer": "B"}},
{{"id": 3, "reasoning": "The correct option is C because...", "answer": "C"}},
]
}}
'''
```

Figure 11: Prompt of the Ref Answer Maker.
```

VLM Pairwise Judge (Video)

```
r'''
You are an expert multimodal AI judge tasked with comparing two video captions.

Given a video and two candidate captions, determine which caption is of higher quality based on fidelity and completeness relative to the
**entire clip**—including how events unfold over time.

# Core principle
Judge CONTENT QUALITY ONLY. Ignore writing style, fluency, verbosity, or length unless they affect factual accuracy. A longer caption is NOT
inherently better.

# Evaluation dimensions

1. **Event and temporal fidelity**: The caption must match what actually happens in the video. Penalize wrong main events, wrong order of events,
confused before/after, or ignoring clear temporal change.

2. **Action, motion, and actors**: Penalize wrong actions, wrong who-did-what, wrong counts of people or objects, or misidentified roles when
the video makes them clear.

3. **Visual grounding**: Objects, attributes, spatial layout, visible text/OCR, and scene setting must align with the video. Hallucinated
objects or attributes are serious errors.

4. **Coverage**: The caption should reflect salient content across the clip when relevant (beginning, middle, end—not only the first frame).
Missing minor background details is acceptable; missing the main event or sequence is not.

5. **Hallucination vs omission**: Penalize hallucinations more heavily than omissions. Generic safe text that avoids real commitments is weak if
the video supports specific, verifiable descriptions.

6. **Specificity**: Prefer specific, grounded descriptions when the video supports them, over vague filler—provided every added detail is
supported.

# Anti-gaming rules
- Do NOT favor a caption because it is longer or more ornate.
- Do NOT penalize reasonable omission of tiny, non-salient background details.
- DO penalize confident statements that contradict the video, even if the rest sounds plausible.
- A broadly correct but generic caption should rank lower than one that is accurate and appropriately specific.

# Tie policy
Use a tie only when both captions are genuinely equivalent in factual precision, temporal coherence, coverage, and grounding, with no meaningful
difference. Do not default to a tie when uncertain—pick the caption that is better supported by evidence across the clip.

Caption 1:
{caption1}

Caption 2:
{caption2}

# Response format
Format your response into exactly two lines:
Reason: <your concise reasoning, comparing both captions against video evidence>
Judgment: <Caption 1 is better> / <Caption 2 is better> / <Tie>'''
```

Figure 12: Prompt of the VLM pairwise judge.

VLM Pointwise Judge (Video)

```
r'''
You are a strict video-caption reward judge. You are given a video and one generated caption. Your job is to score how trustworthy and complete
the caption is with respect to the video.


# Core principle
Judge CONTENT QUALITY ONLY. Ignore writing style, fluency, elegance, verbosity, and length unless they affect factual correctness. A longer
caption is NOT better.


# Important anti-hacking rules
1. Penalize hallucinations more heavily than omissions.
2. Do not reward generic captions that avoid concrete commitments if the video clearly shows identifiable objects, actions, or events.
3. Only reward details that are supported by the video.
4. Strongly penalize wrong temporal order, wrong action, wrong actor, wrong count, wrong scene, or other false factual claims.
5. If the caption adds many unsupported details, the score must stay low even if part of it is correct.


# Internal evaluation process
First, silently decompose the caption into atomic claims and compare them against the video.
Check the beginning, middle, and end when relevant.
For each claim, decide whether it is:
- supported
- contradicted
- not verifiable from the video


Evaluate these 6 dimensions internally:


1. Main event correctness
- Is the central event or activity correctly identified?


2. Object / actor correctness
- Are the important people, animals, objects, and their counts correct?


3. Action / motion correctness
- Are the actions, motion states, and interactions correct?


4. Temporal correctness
- Are event order, change over time, and before/after relations correct?


5. Scene / attribute correctness
- Are setting, salient attributes, and visible text correct?


6. Coverage and specificity
- Does the caption cover the important events and details?
- Do not reward verbose but weakly grounded descriptions.


# Hard caps
- If the caption gets the main event wrong, score <= 2.
- If the caption gets the temporal order badly wrong, score <= 2.
- If the caption contains multiple clear hallucinations, score <= 2.
- If the main event is right but several important details are wrong or omitted, score <= 3.
- If the caption is broadly correct but generic / incomplete, score 3.
- Score 4 only when the caption is accurate on almost all salient visual and temporal content with only minor omissions.
- Score 5 only when the caption is highly accurate, temporally coherent, well-grounded, and covers nearly all important content without
meaningful hallucinations.


# Score scale
1 = Mostly wrong or dominated by hallucination.
2 = Partly related to the video, but contains major factual or temporal errors.
3 = Main event is mostly correct, but the caption is noticeably incomplete, generic, or has some minor mistakes.
4 = Accurate and reasonably complete; only small omissions or very minor errors.
5 = Excellent: precise, temporally coherent, comprehensive, and fully grounded in the video.


# Output format
Output ONLY a valid JSON object:
{{"score": <integer 1-5>}}


Text Caption: {caption}
Your evaluation:
'''
```

Figure 13: Prompt of the VLM pointwise judge.

Offline QA Maker (Video)

```
r'''
# Role
You will act as a professional Visual Data Auditor, specializing in creating comprehensive benchmarks for assessing the quality of video
captions.


# Task
Your primary task is to analyze the provided visual content and design a **comprehensive set of 8 to 15 questions**. These questions will be
used to rigorously evaluate whether a generated video caption is descriptive, accurate, and truly covers all visible details of the video.


# Core Constraint: Objective Observation Only
**CRITICAL:** You must **NOT** ask extended, inferential, or subjective questions.
* **Do NOT ask about internal states:** e.g., "Why is he sad?" (Instead, ask: "Is the person crying?")
* **Do NOT ask about future predictions:** e.g., "What will happen next?"
* **Do NOT ask for external knowledge:** e.g., "Which city is this?" (unless a sign explicitly says it).
* **ONLY ask about what is explicitly evident:** Focus strictly on visible pixels.


# Audit Dimensions (9 Categories)
You MUST generate questions that cover a diverse range of the following 9 categories to ensure maximum coverage:


### --- Visual & Spatial (The "What", "Where", "How") ---
1. **Background (Bkg):** Focus on scene-background details or the setting, ignoring the main subject.
2. **Event (Evt):** Require mention of a specific event, temporal change, or sequence of events.
3. **Instance (Ins):** Emphasise specific entities or compare multiple entities (e.g., counting, presence).
4. **Instance Action (IAct):** Describe the specific motion state or physical action of a designated instance.
5. **Instance Appearance (IApp):** Characterise visual attributes (color, shape, clothing, text/OCR) of a designated instance.
6. **Instance Position (IPos):** Specify the spatial position of an instance within the scene (left, right, center).
7. **Movement (Mov):** Specify camera motion type (e.g., pan, zoom, dolly, static).
8. **Perspective (Per):** Require a particular viewpoint (e.g., bird's-eye view, low-angle, POV).
9. **Region (Reg):** Restrict description/question to a specified image/video region (e.g., "In the top-left corner...").


# Principles of Coverage
1. **Temporal:** Ensure questions cover the **beginning, middle, and end** of the clip.
2. **Spatial:** Do not just focus on the center. Ask about background details and corners.
3. **Randomization:** You must **actively randomize the position of the correct answer** among the options (A, B, C, D).


# Output Format
Please strictly follow the format below. You must respond in a single JSON format.


Example format (The options must be placed in the questions and it is forbidden to form a tag separately.):
```json
{{
"Question_1": "At the very beginning of the video, what object is placed on the far left side of the desk? The options are as follows: A. A red
mug, B. A stack of papers, C. A laptop, D. A potted plant.",
"Answer_1": "D",
"Question_2": "What color is the jacket of the person who enters the room at the 0:15 mark? The options are as follows: A. Black, B. Navy Blue,
C. Bright Yellow, D. Beige.",
"Answer_2": "C",
"Question_3": "According to the visible text on the shop sign, what is the name of the store? The options are as follows: A. Joe's Pizza, B. The
Corner Bakery, C. City Market, D. Fresh Fruits.",
"Answer_3": "B",
"Question_4": "What specific ingredient is visibly placed next to the cutting board? The options are as follows: A. Garlic, B. Chili flakes, C.
Olive oil, D. Salt.",
"Answer_4": "B",
"Question_5": "What visual action happens immediately after the glass falls from the table? The options are as follows: A. The screen goes black,
B. The woman covers her mouth, C. The dog runs away, D. The man laughs.",
"Answer_5": "B",
"Question_6": "Counting only the people sitting on the bench in the background, how many are there? The options are as follows: A. One, B. Two,
C. Three, D. Zero.",
"Answer_6": "B"
}}
'''
```

Figure 14: Prompt of the Offline QA Maker.

Offline QA Solver (Video)

```
r'''
You are a strict text analysis assistant.


**Task:**
You will be provided with a **Video Caption** and a batch of **Multiple-Choice Questions**. For each question, select the correct option based
**ONLY** on the information explicitly stated or clearly implied in the caption.


**Rules:**
1. **Strict Adherence:** Do not use any external knowledge, assumptions, or common sense. Only use the text provided in the caption.
2. For each question, output the letter of the correct option (e.g., "A", "B", "C", etc.).
3. If the caption does not contain enough information to answer a question or support any option, return "None" for that question.


**Caption:**
{caption}


**Questions:**
{questions_json}


**Output:**
Output ONLY a valid JSON object:
{{"answers": [{{"id": 0, "answer": "B"}}, {{"id": 1, "answer": "A"}}, ...]}}
'''
```

Figure 15: Prompt of the Offline QA Solver.
```

Diff Question Maker(Video)

```
r'''# Role
You are an expert in Video Semantic Understanding and Evaluation. Your task is to compare two captions ($Cap1$ and $Cap2$) of the same video and
generate discriminative questions that expose the critical visual differences between them.


# Goal
Produce questions that target meaningful visual differences. Quality over quantity: one sharp question is worth ten trivial ones.


# Input
- $Cap1$: {cap1}
- $Cap2$: {cap2}


# Step-by-Step Reasoning


**Step 1: Identify All Candidate Differences**
List every factual difference between the two captions. Ignore synonyms (e.g., "yelling" vs "shouting") and micro-adjective variations (e.g.,
"light blue" vs "blue").


**Step 2: Rank by Salience (CRITICAL)**
Score each difference on a 1–3 scale and keep only those scoring 3 or 2:
- **3 (Core Narrative)**: Changes the main subject, main action, or overall scene understanding (e.g., person vs animal, running vs standing,
day vs night).
- **2 (Significant Detail)**: Adds or contradicts an important visual element that a viewer would clearly notice (e.g., presence of a key
object).
- **1 (Minor / Peripheral)**: Color nuances, background decorations, generic filler — **discard these**.


**Step 3: Select Salient Differences**
From the remaining candidates (score ≥ 2), keep the salient visual differences. Do NOT pad with low-salience items.


**Step 4: Question Construction & Answer Extraction**
For each selected difference, build one multiple-choice question and determine the answer for each caption.


* **Conflict (Cap1 says A, Cap2 says B)**
* Format: `[Question]? The options are as follows: A. [Cap1 detail], B. [Cap2 detail], C. Neither.`
* `derived_answer_cap1` = "A", `derived_answer_cap2` = "B"


* **Extra_Info (one caption has a major detail the other omits)**
* Format: `[Question]? The options are as follows: A. Yes, B. No.`
* If Cap1 has extra info: `derived_answer_cap1` = "A", `derived_answer_cap2` = "B"
* If Cap2 has extra info: `derived_answer_cap2` = "A", `derived_answer_cap1` = "B"


# Output Guidelines


Output a single JSON object.


**Field Definitions:**
* `type`: "Conflict" or "Extra_Info".
* `salience`: 2 or 3 (the score from Step 2).
* `diff_description`: One sentence summarising the difference.
* `question`: Full question string including options.
* `derived_answer_cap1`: Option letter for Cap1.
* `derived_answer_cap2`: Option letter for Cap2.


# Example (Few-Shot)


**Input:**
Cap1: "A man sprints down a crowded street. A police car is visible in the background."
Cap2: "A boy is walking slowly on a quiet sidewalk."


**Output JSON:**
{{
"questions": [
{{
"type": "Conflict",
"salience": 3,
"diff_description": "Cap1 describes running; Cap2 describes walking.",
"question": "What is the person doing? The options are as follows: A. Sprinting / running fast, B. Walking slowly, C. Neither.",
"derived_answer_cap1": "A",
"derived_answer_cap2": "B"
}},
{{
"type": "Extra_Info",
"salience": 2,
"diff_description": "Cap1 explicitly mentions a police car in the background; Cap2 does not.",
"question": "Is a police car visible in the scene? The options are as follows: A. Yes, B. No.",
"derived_answer_cap1": "A",
"derived_answer_cap2": "B"
}}
]
}}

# Your Task
Process the input provided above and output **ONLY the JSON**, selecting strictly by salience.'''
```

Figure 16: Prompt of the Diff Question Maker.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes].

   Justification: The abstract and introduction state the reward-design problem, the DiffCap-RL framework, and the scope of dense image and video captioning. The main experimental claims are supported by Sec. 4, including main results, reward studies, and ablations.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes].

   Justification: Appendix G discusses the limitation.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [N/A].

Justification: The paper does not present formal theoretical results, theorems, or proofs. The equations in Sec. 3 define the optimization objective and reward components used by the method.

Guidelines:

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

4. **Experimental result reproducibility**

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

Answer: [Yes].

Justification: Sec. 3 specifies the reward construction and optimization procedure, while Sec. 4.1 reports the training data size, base models, judge/verifier models, rollout size, optimizer, learning rate, batch size, response length, and reward hyperparameters. Appendix E provides prompt templates and additional implementation details.

Guidelines:

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

some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [No].

   Justification: At submission time, the paper does not yet provide a public code, model, or dataset URL. We plan to release after acceptance.

   Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes].

   Justification: Sec. 4.1 describes the training corpus, models, optimizer, hyperparameters, baselines, benchmarks, metrics, and evaluation prompts. Appendix B studies the training-data choice and the Diff Scorer weight α.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [No].

   Justification: The paper reports benchmark scores on fixed evaluation sets following the standard protocols of each benchmark, together with human-alignment results and ablations. We do not report error bars or confidence intervals because the primary comparisons are based on deterministic benchmark evaluations rather than repeated random splits, and the main conclusions are supported by consistent gains across multiple benchmarks whose margins are large relative to typical small evaluation fluctuations.

Guidelines:

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

8. **Experiments compute resources**

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

   Answer: [Yes].

   Justification: Section 4.1 reports implementation details.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes].

   Justification: The research is conducted on captioning and reward design using cited existing models, benchmarks, and open-source-based training data. The authors have reviewed the NeurIPS Code of Ethics and are not aware of any deviations.

   Guidelines:

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

10. **Broader impacts**

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

Answer: [Yes].

Justification: Better dense captions may improve accessibility, media understanding, and supervision for multimodal models. Potential risks include misuse in surveillance or misleading media pipelines if deployed without privacy, consent, and bias safeguards; downstream deployments should therefore include appropriate data-governance and safety checks.

Guidelines:

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

11. **Safeguards**

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

Answer: [N/A].

Justification: The paper does not introduce a new high-risk pretrained foundation model, image generator, or newly scraped raw dataset. Released assets will follow the licenses and usage terms of the underlying open-source data and models.

Guidelines:

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

12. **Licenses for existing assets**

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

Answer: [Yes].

Justification: The paper cites the existing models, datasets, benchmarks, and evaluation toolkits used in training and evaluation, including Qwen3-VL, OneThinker, CaptionQA, VDC, VidCap, Capability, and lmms-eval. We will respect and document the corresponding licenses and terms in the released assets.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

Answer: [N/A].

Justification: The paper does not release new assets.

Guidelines:

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

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

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

Answer: [No].

Justification: VideoCapArena uses manual annotations from five PhD researchers participating voluntarily as research collaborators.

Guidelines:

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

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

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

    Answer: [N/A].

    Justification: The work evaluates model-generated captions rather than conducting experiments on human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

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

    Answer: [Yes].

    Justification: LLMs and VLMs are core components of the reward pipeline, including the offline QA maker, vision-free judge, Diff Question Maker, and Ref Answer Maker. Their roles and model choices are described in Sec. 3, Sec. 4.1, and Appendix E.

    Guidelines:

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