VCAP: HYPERGEOMETRIC REWARDS FOR WEAK-TO-STRONG VISUAL CAPTIONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Visual captioning converts visual information into text, requiring both completeness of salient content and correctness of expressed claims. Multimodal large language models (MLLMs) have advanced captioning through scaling and high-quality data, while reinforcement learning (RL) offers a way to refine these factual dimensions. However, directly judging captions with an MLLM can yield coarse quality scores, while visual question answering (VQA) rewards supervise only queried facts. Both routes struggle to provide fine-grained rewards for individual omissions and errors. We propose **VCap**, a **Witness–Adjudicator** fact-comparison reward that uses reference facts as inspection cues and visual evidence to adjudicate omissions and contradictions. References may be incomplete or imperfect; rewarding supported details beyond them enables weak-to-strong learning. A unified hypergeometric model explains how partial factual checks detect both coverage deficits and erroneous assertions under explicit sampling assumptions. Global and local temporal comparisons extend the same reward protocol from images to videos. Training Qwen3-VL-8B with VCap yields the highest aggregate scores among the evaluated models on CapMAS, DecapBench, and VDC. Ablations support the complementary reward components, and human assessment supports more complete captions. RL also outperforms backbone best-of-8 distillation, while the learned captions and policies benefit supervised training and visual question answering.

1 INTRODUCTION

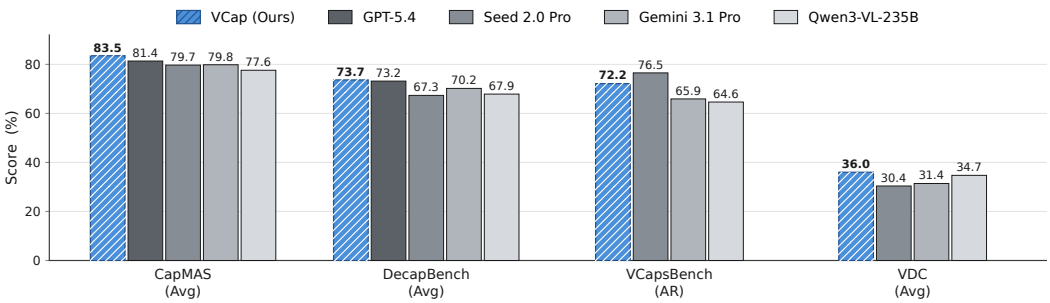


Figure 1: An 8B captioner trained with VCap competes with frontier models on image and video captioning. Tables 2 and 3 report the individual metrics and remaining gaps.

Visual captioning converts information from an image or video into text. Its quality has two central dimensions: **completeness**, preserving salient visual content, and **correctness**, remaining faithful to the visual evidence. Omissions lose useful information; unsupported claims distort it. Both errors affect the captions' value for retrieval, alignment, and model training.

Multimodal large language models (MLLMs) have advanced captioning through vision–language pretraining and increasingly detailed supervision (Li et al., 2023a; Chen et al., 2023; 2024c). Reinforcement learning (RL) offers a further route: score generated responses and optimize their quality

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

beyond imitation of fixed targets (Shao et al., 2024; Shen et al., 2025). For captioning, this makes the reward's ability to assess factual correctness and coverage central to learning.

The main obstacle is obtaining **fine-grained factual rewards**. Two common routes are direct MLLM judging and VQA-based verification. An MLLM judge assigns a quality score to the caption (Xiong et al., 2024), but holistic scores give little indication of which facts are missing or wrong (Huang et al., 2026). CapRL further reports reward hacking driven by judges' preferences for verbosity or brevity (Xing et al., 2025). VQA-based rewards instead measure how well a caption supports answers to image questions (Xing et al., 2025). Their factual coverage is limited by the question pool: adding an unsupported claim or omitting an unqueried detail need not change answer accuracy. Both routes can therefore miss factual quality changes. A useful caption reward must verify individual facts and translate detected omissions and errors into graded supervision.

We introduce **VCap**, a **Witness–Adjudicator** reward built around fact comparison. Comparing reference and candidate captions exposes possible omissions and contradictions; visual evidence resolves disagreements and validates supported details absent from the reference. A frozen multimodal judge grades completeness and correctness by confirmed defects, alongside an auxiliary quality score. References may be incomplete or imperfect; using them as inspection cues allows weak-to-strong learning without restricting the candidate to their factual content.

A unified hypergeometric model explains defect detection: under an idealized fact-slot sampling model, a reference exposes a defect when it probes the corresponding slot. Completeness and correctness share this detection law with different defect sets. Rubrics translate detected omissions and errors into graded scores. Global and local temporal comparisons extend the same reward interface and weights to video. Regenerating references refreshes supervision; inference uses neither references nor a judge.

We train Qwen3-VL-8B-Instruct with GRPO. VCap surpasses the evaluated larger open and proprietary models on the CapMAS, DecapBench, and VDC averages, demonstrating the value of factual reward design for compact captioners. Ablations connect gains to the reward components, human assessment examines caption quality, and distillation and QA transfer establish broader utility.

We summarize our contributions as follows:

- **Fact-comparison rewards.** Reference facts guide inspection, while visual evidence adjudicates completeness and correctness, allowing captions to improve beyond incomplete references.
- **Unified factual detection analysis.** One hypergeometric model describes both coverage-deficit and error detection, explaining how partial references expose the two kinds of defect.
- **Weak-to-strong image and video captioning.** A shared reward protocol trains competitive 8B captioners, with gains supported by ablations, human assessment, distillation, and QA transfer.

2 METHODOLOGY

VCap evaluates completeness and correctness through reference-guided comparison and visual adjudication. We define both dimensions, analyze detection under partial references, and extend the reward to image and video training (Figure 2).

2.1 CAPTIONING AS VISUAL-TO-TEXT INFORMATION CONVERSION

Let x be an image or video and y its textual description. The conversion $x \mapsto y$ selects visual content and expresses it as claims $\Phi(y)$. A conceptual set $\mathcal{F}(x)$ contains the describable facts. Completeness concerns preservation of its salient content; correctness concerns whether the expressed claims are supported by x. Output length, model capacity, and visual ambiguity constrain both dimensions.

Available reference captions y_{ref} expose only part of this content. Supervision is therefore sparse in factual coverage: many scene facts have no explicit reference annotation. A candidate may be correct but incomplete, or detailed but inaccurate. The reward must assess both dimensions while recognizing supported details that its reference does not mention.

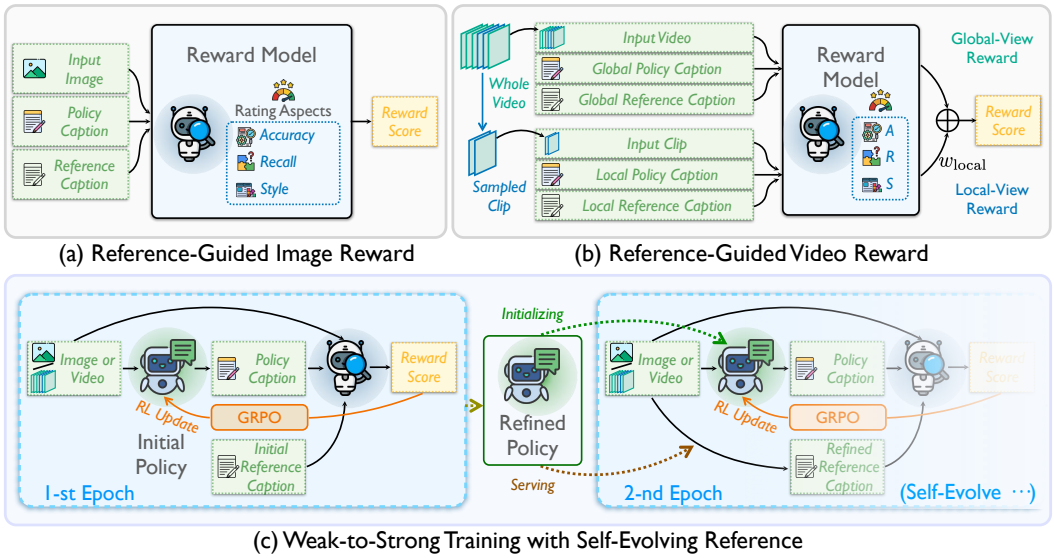


Figure 2: **VCap overview.** (a) Image reward evaluates correctness, completeness, and auxiliary quality (Accuracy, Recall, and Style in the figure). (b) Video combines global and local rewards; local checks use caption-defined segments and the full reference, with $w_{\text{local}} = 0.1$. (c) Both e1 and e2 start from the original backbone; e2 uses references regenerated by e1.

2.2 Fact-Comparison Reward Modeling

Reference-guided comparison, visual adjudication. A frozen Qwen3-VL-32B-Instruct judge takes (x, y_{ref}, y) and returns a structured judgment in one evaluation pass. Three comparison rules address the supervision gap: a visually supported reference detail absent from the candidate is an omission; a disagreement is resolved using the visual input; and a supported candidate detail absent from the reference remains valid. The reference thus supplies inspection cues, while factual validity remains tied to x. The candidate can improve upon its reference by adding supported details and correcting erroneous descriptions, with the rubric turning these factual judgments into reward.

The instructions in Appendix A.6 produce an analysis and three integer scores in $\{0, \ldots, 10\}$. Factual comparison takes place within this prompted judgment, with no explicit exhaustive fact graph. The judge can also inspect claims beyond those mentioned in the reference. The analysis below isolates the reference-guided checks and explains their role in supervising both factual dimensions.

Reward aggregation. The correctness score s_{corr} penalizes unsupported or contradictory claims, and completeness s_{comp} evaluates coverage. An auxiliary score s_{aux} assesses text quality for images and local video segments (fluency, coherence, conciseness, and irrelevant self-evaluation). The global-video instruction instead uses Reasonability in this auxiliary slot. We combine the scores as

$$r(x, y_{\text{ref}}, y) = 0.05 s_{\text{corr}} + 0.04 s_{\text{comp}} + 0.01 s_{\text{aux}}. \tag{1}$$

The weights prioritize factual correctness and completeness and are fixed across image and video runs. The scalar lies in $[0, 1]$ and ranks caption quality; it is not a calibrated correctness probability.

2.3 One Hypergeometric Model for Two Factual Checks

We model sparse reference facts as sampled checks that expose candidate defects. Completeness and correctness differ in which fact slots count as defective; one hypergeometric law characterizes both detection events through this choice of defect set.

Fact universe and sampling assumptions. Fix the visual input and candidate. Let $\mathcal{U}$ be a universe of $N \geq 1$ visual fact slots, each with a true value, such as an object's color or an event's temporal order. Let $C \subseteq \mathcal{U}$ contain the slots whose true facts the candidate expresses, with $c = |C|$, and

$H \subseteq \mathcal{U}$ the slots on which it makes an erroneous assertion, with $e = |H|$. A wrong value belongs to a slot; it does not add a new slot to the universe. The reference supplies correct values for a uniformly sampled set $W \subseteq \mathcal{U}$ of m distinct slots, drawn without replacement and independently of the fixed candidate. Assume error-free adjudication on these probed slots.

A shared detection law. For either check, mark the relevant defect set $D \subseteq \mathcal{U}$, with $d = |D|$. The number of marked slots exposed by the reference follows

$$K_D = |W \cap D| \sim \text{Hypergeom}(N, d, m). \tag{2}$$

Here N is the population size, d the number of marked slots in that population, and m the draw size, with $0 \leq d, m \leq N$. Detection occurs when at least one marked slot is sampled:

$$P_{\text{det}}(N, d, m) = \Pr(K_D \geq 1) = 1 - \frac{\binom{N-d}{m}}{\binom{N}{m}}. \tag{3}$$

Of the $\binom{N}{m}$ reference subsets, $\binom{N-d}{m}$ avoid all marked slots. We set impossible binomial coefficients to zero. Table 1 maps the parameters to both checks; Appendix A.2 derives the law.

Table 1: The same population and reference sample support both checks. Only the marked defect set changes. m counts distinct factual slots, not reference words; e counts erroneous slots, not the number of repeated false statements.

Check	Population size N	Marked set and count d	Draw size m
Completeness	$\lvert\mathcal{U}\rvert$	$D_{\text{comp}} = \mathcal{U} \setminus C, d = N - c$	$\lvert W\rvert$
Correctness	$\lvert\mathcal{U}\rvert$	$D_{\text{corr}} = H, d = e$	$\lvert W\rvert$

Completeness: detect absent correct coverage. A sampled slot in $\mathcal{U} \setminus C$ reveals a true visual fact that the candidate has not correctly expressed. Substituting $d = N - c$ gives

$$P_{\text{det}}^{\text{comp}} = P_{\text{det}}(N, N - c, m) = 1 - \frac{\binom{c}{m}}{\binom{N}{m}}. \tag{4}$$

Its complement is the original reference-coverage event $W \subseteq C$: every sampled true fact is correctly represented. A deficit can be an omission or a wrong value replacing the true one, so this is a test of correct coverage rather than omission alone.

Correctness: detect an erroneous assertion. A sampled slot in H exposes a candidate claim that conflicts with its true value. Substituting $d = e$ gives

$$P_{\text{det}}^{\text{corr}} = P_{\text{det}}(N, e, m) = 1 - \frac{\binom{N-e}{m}}{\binom{N}{m}}. \tag{5}$$

Its complement is the original missed-error event $W \cap H = \varnothing$. An omission alone does not trigger this check. A wrong substitution can trigger both; the events are not assumed independent.

Weak references for strong factual supervision. Under the sampling assumptions, $P_{\text{det}}(N, d, m) \geq 1 - (1 - d/N)^m$ (Appendix A.2). Miss probability decays at least geometrically with m: limited reference coverage can expose defects reliably at a fixed defect fraction d/N. Moreover, for $m \geq 1$, $P_{\text{det}} = 0$ exactly when $d = 0$; the ideal defect-free targets $c = N$ and $e = 0$ do not depend on reference size. These properties connect weak references to strong factual supervision: partial coverage can reveal defects without restricting the candidate to the reference's content. Rubrics convert these detections into graded rewards.

From detection to the implemented reward. The count K_D conveys the extent of detected defects, with $\mathbb{E}[K_D] = md/N$. Rubric engineering translates these checks into graded rewards: the judge lists confirmed omissions and erroneous claims, then adjusts scores according to their number and significance. Fewer omissions merit higher completeness; fewer errors merit higher correctness, with greater weight on salient facts (Appendix A.6). These qualitative criteria produce the integer scores aggregated in Equation (1). The hypergeometric law explains factual detection; the rubric specifies how detected issues affect reward.

Scope. The analysis isolates the reference-guided channel. Natural references have unequal factual coverage, and reference or judge errors affect detection. Appendix A.2 gives the nonuniform extension and explains why stronger detection need not imply stronger marginal training signals. The actual judge can inspect additional visual facts; its overall accuracy and the trained policy's gains require empirical evaluation.

2.4 Image and Video Training

Image training uses Equation (1) directly. Video adds temporal scope to the same factual checks: a description must capture both events across the clip and details within individual segments.

Video extension. The captioner segments its description at natural content changes, including shot transitions. Global evaluation covers the full description. Local evaluation compares segment-relevant facts against the corresponding frames, using the full reference caption:

$$r_{\text{video}} = r_{\text{global}} + 0.1 r_{\text{local}}. \tag{6}$$

Global context supports cross-segment events; local scoring checks the relevant part of the generated description. Both modalities share the reward model and factual score weights, with image and video policies trained separately.

Training and reference regeneration. We use GRPO to optimize the scalar reward (Appendix A.1.3). VCap (e1) uses references generated by the original backbone. VCap (e2) starts again from that backbone and uses references regenerated by e1. Regeneration aims to expose additional supported facts to the judge. Its benefit depends on the coverage and accuracy of the new references; the e1/e2 comparison evaluates the resulting training recipe.

3 Experiments

Our experiments test whether fact-comparison rewards enable weak-to-strong captioning, which components contribute to the gains, and whether the learned capability benefits distillation and QA. Image and video evaluations combine benchmark scores with human judgments.

3.1 Setup

We train image and video policies separately from Qwen3-VL-8B-Instruct (Bai et al., 2025) with GRPO and the rewards in Equations (1) and (6), sharing the three score weights. Both e1 and e2 start from the original backbone, using references from that backbone and e1, respectively. Appendices A.1 and A.6 provide data, configurations, and prompts.

We evaluate image captioning on **CapMAS** (Lee et al., 2025), whose Avg is the mean of CLAIR, Coverage, and Factuality, and **DecapBench** (Ye et al., 2025), whose Avg averages Precision and Recall. Video evaluation uses **VCapsBench** (Zhang et al., 2025b) (AR, IR, CR) and **VDC** (Chai et al., 2025), whose Avg averages five reported aspects. Higher is better except for VCapsBench IR. Baselines include the original backbone, larger open models, proprietary models, and CapRL (Xing et al., 2025) for images, using the same scoring pipeline within each benchmark. Gains denoted by % are absolute differences on the percentage scale.

3.2 Weak-to-Strong Captioning Performance

Tables 2 and 3 support three conclusions about the value of fact-comparison rewards.

Factual rewards enable weak-to-strong captioning. Using references from the original 8B backbone, VCap (e1) already exceeds GPT-5.4 on both image-captioning averages. VCap (e2) extends the gains over the backbone to 5.62% on CapMAS and 4.50% on DecapBench, leading every CapMAS metric and DecapBench Recall and DCScore. Thus, a compact captioner can improve both coverage and factuality beyond its reference generator and compete with larger systems. This supports using reference facts to guide evaluation while allowing the policy to exceed their quality.

Table 2: Image-captioning results on CapMAS and DecapBench.

	CapMAS				DecapBench			
Model	CLAIR	Coverage	Factuality	Avg	Precision	Recall	DCScore	Avg
CapRL	79.65	69.53	63.76	70.98	75.93	50.09	59.47	63.01
Qwen3-VL-8B-Instr	86.19	71.30	76.25	77.91	82.73	55.61	65.70	69.17
Qwen3-VL-235B-Instr	84.44	71.23	77.12	77.60	79.76	55.99	64.99	67.88
Qwen3.5-397B	84.97	73.18	77.02	78.39	78.15	56.21	64.53	67.18
Seed 2.0 Pro	88.03	73.43	77.68	79.71	77.74	56.95	64.89	67.35
Gemini 3.1 Pro	89.19	72.00	78.37	79.85	83.76	56.63	66.46	70.20
GPT-5.4	88.18	73.92	81.99	81.36	86.27	60.08	70.10	73.18
VCap (e1)	89.14	73.95	81.48	81.52	**87.80**	58.83	69.52	73.32
VCap (e2)	**89.99**	**74.18**	**86.42**	**83.53**	86.42	**60.92**	**70.35**	**73.67**

Table 3: Video-captioning results on VCapsBench and VDC.

	VCapsBench			VDC					
Model	AR ↑	IR ↓	CR ↑	Background	Camera	Detailed	Main Object	Short	Avg
Qwen3-VL-8B-Instr	63.28	11.77	71.73	25.63	32.54	37.51	35.28	22.96	30.78
Qwen3-VL-235B-Instr	64.64	12.35	73.75	39.64	32.87	38.57	39.00	**23.56**	34.73
Qwen3.5-397B	65.68	12.05	74.68	37.76	30.53	35.87	36.88	22.96	32.80
Gemini 3.1 Pro	65.91	**11.68**	74.62	36.56	28.54	33.98	35.43	22.57	31.41
Seed 2.0 Pro	**76.53**	13.39	**88.36**	34.91	27.39	33.54	34.55	21.54	30.39
VCap (e1)	71.34	14.01	82.96	40.71	33.22	38.90	39.82	23.42	35.21
VCap (e2)	72.15	13.67	83.57	**41.56**	**34.00**	**40.22**	**40.74**	23.50	**36.01**

The reward protocol generalizes across modalities. With the same factual score weights and an added local temporal check, VCap (e2) improves VDC Avg by 5.23% over the backbone and 1.28% over Qwen3-VL-235B-Instruct, leading four of five VDC aspects. The shared design therefore benefits video as well as image captioning. On VCapsBench, AR/CR also improve, but remain below Seed 2.0 Pro; IR worsens from 11.77 to 13.67, identifying error control as a remaining challenge.

Regenerated references provide further useful supervision. The e1-to-e2 gains are 2.01%, 0.35%, and 0.80% on CapMAS, DecapBench, and VDC Avg. Both policies start from the original backbone, so the improvement comes from the training recipe with refreshed references. DecapBench Recall rises while Precision falls from 87.80 to 86.42: regeneration benefits aggregate quality without improving every dimension. It changes factual coverage, errors, and style jointly; the comparison evaluates this combined intervention. The case study in Appendix A.7 shows added architectural and clothing details alongside conflicting or ambiguous claims, illustrating why added detail still needs verification.

3.3 Human Assessment and Reward Agreement

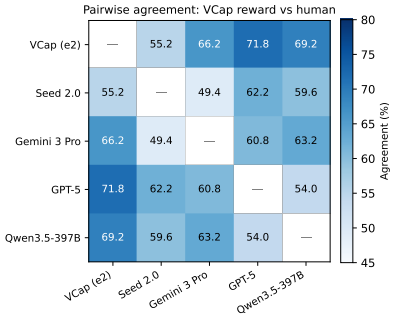
We assess five models on 500 held-out images (Figure 3). Captions are mutually used as references: Qwen3-VL-235B-Instruct proposes missing or inconsistent facts for ordered pairs, an automatic verifier filters proposals, and humans judge the remaining issues following Appendix A.8.

Output quality. VCap has the fewest confirmed omissions (4,137) and the best mean human rank (2.27); its 494 confirmed inconsistencies are second to Seed's 442. Its average caption length is 511 words, close to Seed's 497. These results support detailed output with competitive factual quality. Human review covers only issues proposed by the judge.

Reward agreement. The VCap reward agrees with human rankings in 61.2% of comparisons, averaged over ten model pairs and counting ties as agreement. Together with the output assessment, this connects the reward's ordering to human judgments of factual quality. It measures agreement under this judge-assisted protocol, not independent factual-error detection accuracy.

Model	$\lvert M\rvert$	$\lvert I\rvert$	$\hat{M}$	$\hat{I}$	$\bar{r}_{\mathrm{H}}$	$\tilde{r}_{\mathrm{H}}$	$\bar{r}_{\mathrm{V}}$	$\bar{w}$
VCap (e2)	7,563	**993**	**4,137**	494	**2.27**	**2**	**1.46**	511
Seed 2.0 Pro	**7,318**	1,429	4,289	**442**	2.85	3	2.19	497
Gemini 3.1 Pro	11,906	1,459	6,964	576	2.81	3	3.42	446
GPT-5.4	13,257	1,851	7,997	996	3.58	4	3.71	372
Qwen3.5-397B	15,684	1,367	9,749	593	3.47	4	4.21	351

(a) Per-model statistics



(b) Reward–human agreement

Figure 3: Human assessment on 500 images using judge-proposed factual issues. **(a)** Per-model statistics. $|M|/|I|$: total Judge-proposed missing/inconsistent propositions. $\hat{M}/\hat{I}$: human-confirmed true missing/inconsistent counts. $\bar{r}_{\mathrm{H}}/\tilde{r}_{\mathrm{H}}$: mean and median per-image human rank, where 1 denotes the best model. $\bar{r}_{\mathrm{V}}$: mean per-image rank under the VCap reward. $\bar{w}$: average caption length in words. **(b)** Per-image pairwise agreement (%) between the VCap reward and human ranking, ties counted as agreement under the reported protocol.

Table 4: Image reward ablation. "– modality" removes an input to the judge; "– dimension" sets a score weight to zero. **Bold**/underline mark the best/second-best per column.

	CapMAS				DecapBench			
Setup	CLAIR	Coverage	Factuality	Avg	Precision	Recall	DCScore	Avg
Qwen3-VL-8B-Instr (no RL)	86.19	71.30	76.25	77.91	82.73	55.61	65.71	69.17
VCap (e1) – full reward	**89.14**	73.95	**81.48**	**81.52**	87.80	**58.83**	**69.52**	**73.32**
Modality ablation								
– reference caption	86.45	70.24	76.27	77.65	85.80	53.24	64.82	69.52
– reference image	87.97	73.77	81.37	81.04	86.87	56.45	67.48	71.66
Dimension ablation								
– Correctness	87.16	73.34	76.03	78.84	86.96	58.64	69.09	72.80
– Completeness	88.56	72.94	80.98	80.83	**87.91**	55.26	66.80	71.59
– Text Quality	88.95	**74.18**	81.00	81.38	87.43	58.60	69.17	73.02

3.4 Contributions of the Reward Design

The Witness–Adjudicator design assigns different roles to reference text and visual evidence, and different objectives to completeness and correctness. We test these choices by removing either input or setting one of the three score weights to zero in VCap (e1). Tables 4 and 5 test the same reward components for image and video captioning.

Reference context and visual grounding. Removing the reference lowers CapMAS/DecapBench Avg by 3.87%/3.80%; removing the image lowers them by 0.48%/1.66%. The larger reference effect supports using textual facts as inspection cues, while the image provides additional value even with those cues present. These complementary benefits support the Witness–Adjudicator design, although resolving deliberately corrupted references requires a separate test.

Correctness and completeness. Removing completeness reduces Coverage/Recall by 1.01%/3.57%; removing correctness reduces Factuality/Precision by 5.45%/0.84%. Each score therefore benefits its intended factual dimension. Removing text quality lowers CapMAS/DecapBench Avg by only 0.14%/0.30%. The dominant gains come from the two factual objectives, matching the information-conversion account of caption quality.

Table 5: Video reward ablation. Removing the reference, a score dimension, or the local term tests its contribution (Appendix A.5). **Bold**/underline mark the best/second-best per column.

	VCapsBench			VDC					
Setup	AR↑	IR↓	CR↑	Background	Camera	Detailed	Main Object	Short	Avg
Qwen3-VL-8B-Instr (no RL)	63.28	**11.77**	71.73	25.63	32.54	37.51	35.28	22.96	30.78
VCap (e1) – full reward	**71.34**	14.01	**82.96**	**40.71**	**33.22**	**38.90**	**39.82**	23.42	**35.21**
– reference caption	65.75	12.94	76.19	37.51	32.29	37.16	39.35	23.54	33.97
– Correctness	69.14	13.21	79.66	**40.71**	32.42	38.57	39.59	23.76	35.01
– Completeness	69.41	12.16	79.02	38.75	32.72	38.21	39.53	23.56	34.55
– Text Quality	70.73	12.53	80.86	40.50	33.01	38.73	39.72	23.47	35.09
– local term	70.94	12.81	81.25	39.49	32.62	36.35	39.44	**24.07**	34.39

Reference and local checks support video captioning. Table 5 supports reference-guided factual evaluation with complementary global and local checks. Removing the reference lowers VCaps-Bench AR/CR by 5.59%/6.77%; removing completeness or the local term reduces VDC Avg by 0.66%/0.82%, respectively. Removing correctness or text quality has smaller effects on VDC Avg. The full reward's higher IR indicates that better coverage still leaves room to improve error control.

3.5 Training Compact Captioners: RL and Distillation

A compact policy can learn through reward optimization or imitation of selected outputs. Table 6 compares VCap with three SFT runs initialized from Qwen3-VL-8B-Instruct. For each input, the VCap reward selects the best of eight captions generated by the backbone, e1, or e2. This tests a fixed best-of-8 training recipe, leaving broader sampling budgets unevaluated.

Table 6: Best-of-8 distillation. Students start from Qwen3-VL-8B-Instruct and imitate reward-selected captions from each teacher. **Bold**/underline mark the best/second-best per column.

	CapMAS				DecapBench			
Model	CLAIR	Coverage	Factuality	Avg	Precision	Recall	DCScore	Avg
Qwen3-VL-8B-Instr	86.19	71.30	76.25	77.91	82.73	55.61	65.70	69.17
+ Self-distill	84.84	71.09	77.02	77.65	80.98	54.95	64.86	67.97
VCap (e1)	89.14	73.95	81.48	81.52	**87.80**	58.83	69.52	73.32
+ Distill from VCap (e1)	88.62	73.00	84.37	82.00	86.50	58.56	68.71	72.53
VCap (e2)	**89.99**	**74.18**	**86.42**	**83.53**	86.42	**60.92**	**70.35**	**73.67**
+ Distill from VCap (e2)	89.10	73.71	86.29	83.03	86.22	59.75	69.44	72.99

Backbone self-distillation falls below the original model on both averages, whereas VCap RL improves both. Thus, selecting and imitating the backbone's existing outputs does not recover the RL gains in this setting. Distillation from VCap teachers does transfer substantial gains: the e2 student reaches 83.03/72.99, and the e1 student even exceeds its teacher on CapMAS Avg (82.00 versus 81.52). VCap therefore improves the captioner and produces useful training targets. These conclusions apply to the evaluated teachers and best-of-8 procedure.

3.6 Transfer Beyond Captioning

We evaluate the captioning checkpoints on QA without task-specific fine-tuning (Table 7). Image QA uses Prism (Qiao et al., 2024) on AI2D, MMStar, RWQA (RealWorldQA), and V⋆; video QA directly evaluates MMVU, MLVU, VMMMU (VideoMMMU), and LVBench. All models share the evaluation pipeline within each setting.

Captioning rewards yield broader task benefits. VCap (e2) improves the image/video QA averages by 1.93%/1.54%, with gains on all four image datasets and three of four video datasets. Its

Table 7: QA transfer of captioning checkpoints without QA-specific fine-tuning. Image QA uses Prism; video QA evaluates the policies directly. Each Avg is the mean of four benchmark scores.

	Image QA (Prism)					Video QA				
Model	AI2D	MMStar	RWQA	V⋆	Avg	MMVU	MLVU	VMMMU	LVBench	Avg
CapRL	83.71	66.87	63.27	52.36	66.56	–	–	–	–	–
Qwen3-VL-8B-Instr	81.83	65.20	60.78	50.79	64.65	60.80	74.20	57.00	45.64	59.41
VCap (e1)	83.19	66.53	60.92	48.17	64.70	62.50	74.43	58.44	46.16	60.38
VCap (e2)	83.81	67.13	62.48	52.88	66.58	63.10	73.92	60.44	46.35	60.95
Qwen3-VL-235B-Instr	85.40	69.00	62.88	52.88	67.54	70.30	80.59	72.67	54.55	69.53
Qwen3-VL-235B-Think	86.04	70.27	66.01	47.64	67.49	74.30	78.98	75.78	50.87	69.98

image average (66.58) is comparable to CapRL (66.56). These improvements show that the learned capability is useful beyond caption scores, supporting captioning as a training task for visual understanding. Transfer remains uneven: MLVU declines from 74.20 to 73.92, and the 235B models retain a substantial video-QA advantage. The results establish downstream utility without isolating visual perception as the sole mechanism.

4 Related Work

Factual rewards for captioning. Existing approaches improve caption supervision through complementary routes: CapRL (Xing et al., 2025) uses VQA-based rewards, LLaVA-Critic (Xiong et al., 2024) learns multimodal evaluation, and MDPO/V-DPO (Wang et al., 2024a; Xie et al., 2024) optimize preference objectives. Fact localization, precision–recall criteria, and rubric scoring further refine evaluation (Wang et al., 2025b; Zhang et al., 2025a; Huang et al., 2026; Ye et al., 2025). VCap pursues fine-grained factual supervision through reference-guided comparison and visual adjudication. Reference facts expose possible defects, visual evidence adjudicates factual support, and correct candidate details beyond the reference remain eligible for reward.

Detection analysis and supervision reuse. Synthetic recaptioning improves training targets (Yu et al., 2024a), while VCap uses references to guide factual inspection and supports visually grounded content beyond them. The Witness–Adjudicator protocol connects these information roles to a shared image/video reward interface. One hypergeometric model describes both factual checks through different defect sets; temporal assessment changes the scope of a check, and reference regeneration changes the probe pool. The analysis characterizes reference-guided detection under explicit assumptions. Input and score ablations evaluate the implemented reward, while matched judge comparisons are needed to distinguish closely related protocols. Appendices A.3 and A.4 discuss the broader literature.

5 Conclusion

VCap supervises visual-to-text information conversion through completeness and correctness. Reference-guided fact comparison and visual adjudication let captions improve beyond imperfect references, with hypergeometric analysis explaining how partial checks expose defects. The shared image/video reward trains 8B captioners that lead the evaluated models on CapMAS, DecapBench, and VDC, supported by ablations, human assessment, distillation, and QA transfer. The Keye-VL 2.0 report describes pretraining captions from a VCap-trained model (Kwai Keye Team, 2026).

Limitations. Fact detection in realistic visual descriptions is more complex than the idealized sampling process: which facts are exposed and how they are judged depend jointly on the reference, visual evidence, and candidate caption. Our analysis characterizes a tractable part of this process, while its correspondence with practical reward behavior and weak-to-strong learning requires deeper investigation. We are extending the analysis and refining the experimental design. Further large-scale ablation studies are underway to quantify the method's contribution.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

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. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=tTDUrseRRU.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pp. 3558–3568. Computer Vision Foundation / IEEE, 2021. doi: 10.1109/CVPR46437.2021.00356. URL https://openaccess.thecvf.com/content/CVPR2021/html/Changpinyo_Conceptual_12M_Pushing_Web-Scale_Image-Text_Pre-Training_To_Recognize_Long-Tail_Visual_CVPR_2021_paper.html.

Guiming Hardy Chen, Shunian Chen, Ruifei Zhang, Junying Chen, Xiangbo Wu, Zhiyi Zhang, Zhihong Chen, Jianquan Li, Xiang Wan, and Benyou Wang. Allava: Harnessing gpt4v-synthesized data for lite vision-language models. *ArXiv preprint*, abs/2402.11684, 2024a. URL https://arxiv.org/abs/2402.11684.

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. *ArXiv preprint*, abs/2311.12793, 2023. URL https://arxiv.org/abs/2311.12793.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024b.

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Lin Bin, Zhenyu Tang, Li Yuan, Yu Qiao, Dahua Lin, Feng Zhao, and Jiaqi Wang. Sharegpt4video: Improving video understanding and generation with better captions. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024c. URL http://papers.nips.cc/paper_files/paper/2024/hash/22a7476e4fd36818777c47e666f61a41-Abstract-Datasets_and_Benchmarks_Track.html.

Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, and Sergey Tulyakov. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 13320–13331. IEEE, 2024d. doi: 10.1109/CVPR52733.2024.01265. URL https://doi.org/10.1109/CVPR52733.2024.01265.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *ArXiv preprint*, abs/1504.00325, 2015. URL https://arxiv.org/abs/1504.00325.

Xinlong Chen, Yuanxing Zhang, Yushuo Guan, Weihong Lin, Zekun Wang, Bohan Zeng, Yang Shi, Sihan Yang, Qiang Liu, Pengfei Wan, Liang Wang, and Tieniu Tan. Vidbridge-R1: Bridging QA and captioning for RL-based video understanding models with intermediate proxy tasks. *ArXiv preprint*, abs/2506.09079, 2025. URL https://arxiv.org/abs/2506.09079.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, et al. Deepseek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *ArXiv preprint*, abs/2501.12948, 2025. URL https://arxiv.org/abs/2501.12948.

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

Xueqing Deng, Qihang Yu, Peng Wang, Xiaohui Shen, and Liang-Chieh Chen. COCONut: Modernizing COCO segmentation. *arXiv preprint arXiv:2404.08639*, 2024. URL https://arxiv.org/abs/2404.08639.

Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. Improving CLIP training with language rewrites. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/6fa4d985e7c434002fb6289ab9b2d654-Abstract-Conference.html.

Roopal Garg, Andrea Burns, Burcu Karagol Ayan, Yonatan Bitton, Ceslee Montgomery, Yasumasa Onoe, Andrew Bunner, Ranjay Krishna, Jason Michael Baldridge, and Radu Soricut. ImageInWords: Unlocking hyper-detailed image descriptions. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 93–127, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.6. URL https://aclanthology.org/2024.emnlp-main.6/.

John Gkountouras and Ivan Titov. Clarification as supervision: Reinforcement learning for vision-language interfaces. *ArXiv preprint*, abs/2509.26594, 2025. URL https://arxiv.org/abs/2509.26594.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 14375–14385. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01363. URL https://doi.org/10.1109/CVPR52733.2024.01363.

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. URL https://arxiv.org/abs/2501.13826.

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

Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. FaithScore: Fine-grained evaluations of hallucinations in large vision-language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 5042–5063, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.290. URL https://aclanthology.org/2024.findings-emnlp.290/.

Xuan Ju, Yiming Gao, Zhaoyang Zhang, Ziyang Yuan, Xintao Wang, Ailing Zeng, Yu Xiong, Qiang Xu, and Ying Shan. Miradata: A large-scale video dataset with long durations

and structured captions. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/57f6683e550eb067936c9e9f0bcb8e31-Abstract-Datasets_and_Benchmarks_Track.html.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. A diagram is worth a dozen images. *arXiv preprint arXiv:1603.07396*, 2016.

Kwai Keye Team. Kwai Keye-VL-2.0 technical report. *ArXiv preprint*, abs/2606.10651, 2026. URL https://arxiv.org/abs/2606.10651.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. *arXiv preprint arXiv:2309.06180*, 2023. URL https://arxiv.org/abs/2309.06180.

Saehyung Lee, Seunghyun Yoon, Trung Bui, Jing Shi, and Sungroh Yoon. Toward robust hyper-detailed image captioning: A multiagent approach and dual evaluation metrics for factuality and coverage. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, Proceedings of Machine Learning Research. PMLR / OpenReview.net, 2025. URL https://proceedings.mlr.press/v267/lee25aj.html.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12888–12900. PMLR, 2022. URL https://proceedings.mlr.press/v162/li22n.html.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023a. URL https://proceedings.mlr.press/v202/li23q.html.

Xiaotong Li, Fan Zhang, Haiwen Diao, Yueze Wang, Xinlong Wang, and Lingyu Duan. Densefusion-1m: Merging vision experts for comprehensive multimodal perception. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/20ffc2b42c7de4a1960cfdadf305bbe2-Abstract-Datasets_and_Benchmarks_Track.html.

Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-R1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *ArXiv preprint*, abs/2504.06958, 2025. URL https://arxiv.org/abs/2504.06958.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305, Singapore, 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.20. URL https://aclanthology.org/2023.emnlp-main.20/.

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

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Fan Lu, Wei Wu, Kecheng Zheng, Shuailei Ma, Biao Gong, Jiawei Liu, Wei Zhai, Yang Cao, Yujun Shen, and Zheng-Jun Zha. Benchmarking large vision-language models via directed scene graph for comprehensive image captioning. *ArXiv preprint*, abs/2412.08614, 2024. URL https://arxiv.org/abs/2412.08614.

Xingyu Lu, Jinpeng Wang, YiFan Zhang, Shijie Ma, Xiao Hu, Tianke Zhang, Haonan fan, Kaiyu Jiang, Changyi Liu, Kaiyu Tang, Bin Wen, Fan Yang, Tingting Gao, Han Li, and Chun Yuan. Contextrl: Enhancing mllm's knowledge discovery efficiency with context-augmented rl. *ArXiv preprint*, abs/2602.22623, 2026. URL https://arxiv.org/abs/2602.22623.

Desen Meng, Rui Huang, Zhilin Dai, Xinhao Li, Yifan Xu, Jun Zhang, Zhenpeng Huang, Meng Zhang, Lingshu Zhang, Yi Liu, and Limin Wang. Videocap-R1: Enhancing MLLMs for video captioning via structured thinking. *ArXiv preprint*, abs/2506.01725, 2025. URL https://arxiv.org/abs/2506.01725.

Kepan Nan, Rui Xie, Penghao Zhou, Tiehan Fan, Zhenheng Yang, Zhijie Chen, Xiang Li, Jian Yang, and Ying Tai. Openvid-1M: A large-scale high-quality dataset for text-to-video generation. *ArXiv preprint*, abs/2407.02371, 2024. URL https://arxiv.org/abs/2407.02371.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F. Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/b1efde53be364a73914f58805a001731-Abstract-Conference.html.

Yuxuan Qiao, Haodong Duan, Xinyu Fang, Junming Yang, Lin Chen, Songyang Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. Prism: A framework for decoupling and assessing the capabilities of vlms. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/cac9e747a1d480c78312226959566cef-Abstract-Conference.html.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. ZeRO: Memory optimizations toward training trillion parameter models. *arXiv preprint arXiv:1910.02054*, 2019. URL https://arxiv.org/abs/1910.02054.

Ruchit Rawal, Reza Shirkavand, Heng Huang, Gowthami Somepalli, and Tom Goldstein. ARGUS: Hallucination and omission evaluation in video-LLMs. *ArXiv preprint*, abs/2506.07371, 2025. URL https://arxiv.org/abs/2506.07371.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: an open large-scale dataset for training next generation image-text models. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/

2022/hash/a1859debfb3b59d094f3504d5ebb6c25-Abstract-Datasets_and_Benchmarks.html.

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

Haozhan Shen, Peng Liu, Jingcheng Li, Chunxin Fang, Yibo Ma, Jiajia Liao, Qiaoli Shen, Zilun Zhang, Kangjia Zhao, Qianqian Zhang, Ruochen Xu, and Tiancheng Zhao. Vlm-R1: A stable and generalizable R1-style large vision-language model. *ArXiv preprint*, abs/2504.07615, 2025. URL https://arxiv.org/abs/2504.07615.

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

Yanpeng Sun, Jing Hao, Ke Zhu, Jiang-Jiang Liu, Yuxiang Zhao, Xiaofan Li, Na Zhao, Zechao Li, and Jingdong Wang. Enhancing descriptive captions with visual attributes for multimodal perception. *ArXiv preprint*, abs/2412.14233, 2024a. URL https://arxiv.org/abs/2412.14233.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. Aligning large multimodal models with factually augmented RLHF. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 13088–13110, Bangkok, Thailand, 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.775. URL https://aclanthology.org/2024.findings-acl.775/.

Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. video-SALMONN 2: Caption-enhanced audio-visual large language models. *ArXiv preprint*, abs/2506.15220, 2025. URL https://arxiv.org/abs/2506.15220.

Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. YFCC100M: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

Jack Urbanek, Florian Bordes, Pietro Astolfi, Mary Williamson, Vasu Sharma, and Adriana Romero-Soriano. A picture is worth more than 77 text tokens: Evaluating clip-style models on dense captions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 26690–26699. IEEE, 2024. doi: 10.1109/CVPR52733.2024.02521. URL https://doi.org/10.1109/CVPR52733.2024.02521.

Fei Wang, Wenxuan Zhou, James Y. Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. mDPO: Conditional preference optimization for multimodal large language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8078–8088, Miami, Florida, USA, 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.460. URL https://aclanthology.org/2024.emnlp-main.460/.

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

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. AMBER: An LLM-free multi-dimensional benchmark for MLLMs hallucination evaluation. *ArXiv preprint*, abs/2311.07397, 2023. URL https://arxiv.org/abs/2311.07397.

Qiang Wang, Xinyuan Gao, SongLin Dong, Jizhou Han, Jiangyang Li, Yuhang He, and Yihong Gong. Vdc-agent: When video detailed captioners evolve themselves via agentic self-reflection. *arXiv preprint arXiv:2511.19436v1*, 2025a. URL https://arxiv.org/abs/2511.19436v1.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, et al. LVBench: An extreme long video understanding benchmark. *arXiv preprint arXiv:2406.08035*, 2024c. URL https://arxiv.org/abs/2406.08035.

Xiyao Wang, Zhengyuan Yang, Chao Feng, Yongyuan Liang, Yuhang Zhou, Xiaoyu Liu, Ziyi Zang, Ming Li, Chung-Ching Lin, Kevin Lin, Linjie Li, Furong Huang, and Lijuan Wang. Vicrit: A verifiable reinforcement learning proxy task for visual perception in VLMs. *ArXiv preprint*, abs/2506.10128, 2025b. URL https://arxiv.org/abs/2506.10128.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal LLMs. *arXiv preprint arXiv:2312.14135*, 2023. URL https://arxiv.org/abs/2312.14135.

xAI. RealWorldQA, 2024. URL https://huggingface.co/datasets/xai-org/RealworldQA.

Wenyi Xiao, Ziwei Huang, Leilei Gan, Wanggui He, Haoyuan Li, Zhelun Yu, Fangxun Shu, Hao Jiang, and Linchao Zhu. Detecting and mitigating hallucination in large vision language models via fine-grained AI feedback. In Toby Walsh, Julie Shah, and Zico Kolter (eds.), *Thirty-Ninth AAAI Conference on Artificial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025*, pp. 25543–25551. AAAI Press, 2025. doi: 10.1609/AAAI.V39I24.34744. URL https://doi.org/10.1609/aaai.v39i24.34744.

Yuxi Xie, Guanzhen Li, Xiao Xu, and Min-Yen Kan. V-DPO: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 13258–13273, Miami, Florida, USA, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.775. URL https://aclanthology.org/2024.findings-emnlp.775/.

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

Tianyi Xiong, Xiyao Wang, Dong Guo, Qinghao Ye, Haoqi Fan, Quanquan Gu, Heng Huang, and Chunyuan Li. LLaVA-Critic: Learning to evaluate multimodal models. *ArXiv preprint*, abs/2410.02712, 2024. URL https://arxiv.org/abs/2410.02712.

Dongjie Yang, Suyuan Huang, Chengqiang Lu, Xiaodong Han, Haoxin Zhang, Yan Gao, Yao Hu, and Hai Zhao. Vript: A video is worth thousands of words. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/6903a5aaece71b76623245fc6e32f01b-Abstract-Datasets_and_Benchmarks_Track.html.

Qinghao Ye, Xianhan Zeng, Fu Li, Chunyuan Li, and Haoqi Fan. Painting with words: Elevating detailed image captioning with benchmark and alignment learning. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=636M0nNbPs.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. doi: 10.1162/tacl_a_00166. URL https://aclanthology.org/Q14-1006/.

Qiying Yu, Quan Sun, Xiaosong Zhang, Yufeng Cui, Fan Zhang, Yue Cao, Xinlong Wang, and Jingjing Liu. Capsfusion: Rethinking image-text data at scale. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 14022–14032. IEEE, 2024a. doi: 10.1109/CVPR52733.2024.01330. URL https://doi.org/10.1109/CVPR52733.2024.01330.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, and Maosong Sun. RLHF-V: towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pp. 13807–13816. IEEE, 2024b. doi: 10.1109/CVPR52733.2024.01310. URL https://doi.org/10.1109/CVPR52733.2024.01310.

Tianyu Yu, Haoye Zhang, Qiming Li, Qixin Xu, Yuan Yao, Da Chen, Xiaoman Lu, Ganqu Cui, Yunkai Dang, Taiwen He, Xiaocheng Feng, Jun Song, Bo Zheng, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. RLAIF-V: open-source AI feedback leads to super GPT-4V trustworthiness. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 19985–19995. Computer Vision Foundation / IEEE, 2025. doi: 10.1109/CVPR52734.2025.01861. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yu_RLAIF-V_Open-Source_AI_Feedback_Leads_to_Super_GPT-4V_Trustworthiness_CVPR_2025_paper.html.

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*, abs/2501.07888, 2025. URL https://arxiv.org/abs/2501.07888.

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

Shi-Xue Zhang, Hongfa Wang, Duojun Huang, Xin Li, Xiaobin Zhu, and Xu-Cheng Yin. VCaps-Bench: A large-scale fine-grained benchmark for video caption quality evaluation. *arXiv preprint arXiv:2505.23484*, 2025b. URL https://arxiv.org/abs/2505.23484.

Yilun Zhao, Lujing Xie, Haowei Zhang, Guo Gan, Yitao Long, Zhiyuan Hu, Tongyan Hu, Weiyuan Chen, Chuhan Li, Junyang Song, et al. MMVU: Measuring expert-level multi-discipline video understanding. *arXiv preprint arXiv:2501.12380*, 2025a. URL https://arxiv.org/abs/2501.12380.

Yuze Zhao, Jintao Huang, Jinghan Hu, Xingjun Wang, Yunlin Mao, Daoze Zhang, Zeyinzi Jiang, Zhikai Wu, Baole Ai, Ang Wang, et al. Swift: a scalable lightweight infrastructure for fine-tuning. In *AAAI*, volume 39, pp. 29733–29735, 2025b.

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

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. MLVU: Benchmarking multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2024. URL https://arxiv.org/abs/2406.04264.

LLM USAGE DECLARATION

LLMs were used solely to check grammar and improve the clarity of the manuscript. They did not contribute to the conception of the research or the generation of its substantive content. The authors assume full responsibility for the entire paper, including any language revisions assisted by LLMs. We confirm that this use adheres to applicable ethical standards and does not involve plagiarism or other forms of scientific misconduct.

A APPENDIX

APPENDIX CONTENTS

A.1 Experimental Setup Details 17
A.2 Details of the Unified Fact-Detection Analysis 18
A.3 Related Work 20
A.4 Reference Roles Across Reward Families 22
A.5 Video-Captioning Reward-System Ablation 22
A.6 Instructions for VCap Reward Modeling 22
A.7 Case Study: Reference, VCap (e1), and VCap (e2) 24
A.8 Human Annotation Protocol 27

A.1 EXPERIMENTAL SETUP DETAILS

A.1.1 TRAINING DATA

Image captioning. We use ∼29k images from COCONut (Deng et al., 2024), a COCO extension with dense panoptic annotations. The initial reference pool $\mathcal{D}_{\text{ref}}(x)$ is generated by QWEN3-VL-8B-INSTRUCT; VCap (e1) regenerates the pool for e2 on the same images under identical decoding settings.

Video captioning. We use ∼13k long-form videos collected from YouTube under a topic distribution that mirrors typical long-video benchmarks (vlogs, documentary, sports, instructional, tutorials, and scripted scenes). Frames are uniformly subsampled when needed to meet the visual input budget. The initial full-video references are produced by QWEN3-VL-8B-INSTRUCT, and e1 regenerates them for e2. The captioner organizes its description around natural content or shot changes. Local evaluation in Equation (6) uses these output-defined segments and the full-video reference, comparing only the facts relevant to the segment under evaluation.

Evaluation resources. Image QA uses AI2D (Kembhavi et al., 2016), MMStar (Chen et al., 2024b), RealWorldQA (xAI, 2024), and V⋆ (Wu & Xie, 2023) through Prism (Qiao et al., 2024). Video QA uses MMVU (Zhao et al., 2025a), MLVU (Zhou et al., 2024), VideoMMMU (Hu et al., 2025), and LVBench (Wang et al., 2024c). RWQA and VMMMU in Table 7 denote RealWorldQA and VideoMMMU, respectively.

A.1.2 VCAP REWARD HYPERPARAMETERS

The image reward is Equation (1), with weights $w_{\text{corr}} = 0.05$, $w_{\text{comp}} = 0.04$, and $w_{\text{aux}} = 0.01$. The auxiliary slot uses Text Quality for images and local video segments, and Reasonability for the global video pass, as specified by the prompts. The scores lie in $\{0, \ldots, 10\}$, so the weighted image reward lies in $[0, 1]$. Video uses the same score weights and adds a local evaluation with $w_{\text{local}} = 0.1$ as in Equation (6). The local term emphasizes a caption-defined segment while the global term retains the full-video context. These weights are empirical design choices. Because factual scores can trade off under length and capacity constraints, the idealized witness analysis does not imply that changing the weights preserves the actual policy optimum.

A.1.3 GRPO ALGORITHM

We optimize a captioner $\pi_\theta(y \mid x)$ against the sentence-level scalar reward $r(x, y_{\text{ref}}, y)$ of Equation (1) (or Equation (6) for video), where y_{ref} is a reference caption drawn from a pool $\mathcal{D}_{\text{ref}}(x)$ associated with x. The reference is auxiliary evaluation context and may be imperfect. Each training step samples a pair (x, y_{ref}), draws a group of G candidate captions $\{y^{(i)}\}_{i=1}^{G} \sim \pi_{\theta_{\text{old}}}(\cdot \mid x)$ from a frozen behavior policy, scores each candidate with $r(x, y_{\text{ref}}, y^{(i)})$, and updates π_θ on the resulting group.

We use GRPO (Shao et al., 2024) as the on-policy RL algorithm, which avoids a learned value baseline by computing per-sample advantages within the on-policy group:

$$\hat{A}^{(i)} \;=\; \frac{r^{(i)} - \mu_r}{\sigma_r}, \quad r^{(i)} \;=\; r\big(x, y_{\text{ref}}, y^{(i)}\big), \quad \mu_r = \text{mean}\big(\{r^{(j)}\}_{j=1}^{G}\big), \quad \sigma_r = \text{std}\big(\{r^{(j)}\}_{j=1}^{G}\big), \tag{7}$$

for groups with nonzero reward variance. GRPO maximizes the clipped surrogate objective with a KL anchor on a frozen reference policy π_{ref}:

$$\mathcal{J}_{\text{GRPO}}(\theta) = \mathbb{E}\Bigg[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|y^{(i)}|}\sum_{t=1}^{|y^{(i)}|}\Big(\min\Big(\rho_t^{(i)}\hat{A}^{(i)},\ \text{clip}(\rho_t^{(i)}, 1{-}\varepsilon, 1{+}\varepsilon)\hat{A}^{(i)}\Big) - \beta D_{\text{KL}}\big(\pi_\theta \| \pi_{\text{ref}}\big)\Big)\Bigg]. \tag{8}$$

where $\rho_t^{(i)} = \pi_\theta(y_t^{(i)} \mid x, y_{<t}^{(i)}) / \pi_{\theta_{\text{old}}}(y_t^{(i)} \mid x, y_{<t}^{(i)})$ is the per-token importance ratio, ε the clipping threshold, and β the KL weight. The scalar reward can be supplied to other reward-based optimizers, but our empirical results use GRPO. The witness analysis concerns idealized reference probes and supplies no optimizer-specific convergence guarantee.

A.1.4 TRAINING CONFIGURATION

All experiments are run on a single 8×NVIDIA H800 node using the ms-swift (Zhao et al., 2025b) with vLLM (Kwon et al., 2023) as the rollout sampler and DeepSpeed ZeRO-3 (Rajbhandari et al., 2019) for parameter and optimizer-state sharding. Both the vision encoder and the vision–language aligner are frozen throughout RL training, so only the language-model parameters of QWEN3-VL-8B-INSTRUCT receive gradients. Table 8 lists the optimization, sampling, and rollout hyperparameters used by the image and video runs. Both runs use AdamW (Loshchilov & Hutter, 2017).

Visual encoding budget. For image captioning we cap the per-image pixel budget at `max_pixels` = 6,422,528, which corresponds to up to ∼8,192 visual tokens per image under the Qwen3-VL patch tokenizer. For video captioning we sample at most 200 frames per video, uniformly subsampling longer sequences. The 32K prompt budget includes both visual and text tokens.

A.2 DETAILS OF THE UNIFIED FACT-DETECTION ANALYSIS

Population, marked slots, and sample. The population $\mathcal{U}$ contains N modeled visual fact slots with well-defined true values. The candidate and its sets C and H are fixed before drawing the reference slots W. The sample consists of m distinct reference slots drawn uniformly without replacement from $\mathcal{U}$. In both applications the population and sample are identical: completeness marks $D_{\text{comp}} = \mathcal{U} \setminus C$, of size $N - c$, while correctness marks $D_{\text{corr}} = H$, of size e. Erroneous values and repeated mentions do not create additional population elements or draws. Claims outside the chosen fact universe and ambiguous true values are outside this idealized model.

For a fixed marked set D of size d, exactly $\binom{d}{k}\binom{N-d}{m-k}$ samples expose k marked slots. Consequently,

$$\Pr(K_D = k) = \frac{\binom{d}{k}\binom{N-d}{m-k}}{\binom{N}{m}}, \qquad \max(0, m-(N-d)) \le k \le \min(d, m). \tag{9}$$

Table 8: Training hyperparameters for image-captioning and video-captioning runs.

Hyperparameter	Image run	Video run
Hardware	8×H800	8×H800
Framework	ms-swift	ms-swift
Rollout sampler	vLLM	vLLM
Distributed strategy	DeepSpeed ZeRO-3	DeepSpeed ZeRO-3
Precision	bfloat16	bfloat16
Attention	FlashAttention	FlashAttention
Optimizer	AdamW	AdamW
Learning rate	5×10^{-7}	5×10^{-7}
LR scheduler	cosine w/ min-LR	cosine w/ min-LR
Min-LR ratio	0.1	0.1
Warmup ratio	0.01	0.003
Epochs	1	1
Per-device batch size	2	1
Gradient accumulation steps	16	32
Effective batch size	256	256
Group size G (`num_generations`)	8	8
Sampling temperature	1.0	1.0
Top-p	0.9	0.9
Top-k	50	50
Repetition penalty	1.05	1.05
KL coefficient β	1×10^{-3}	1×10^{-4}
Max prompt length	14,000 tokens	32,000 tokens
Max completion length	4,096 tokens	6,300 tokens

The parameters obey $N \geq 1$ and $0 \leq d, m \leq N$. Evaluating at $k = 0$ gives the missed-detection probability, and taking its complement yields Equation (3). If $m = 0$ or $d = 0$, detection has probability zero; if $m > N - d$, it has probability one. The first condition describes an empty check or no marked defects, while the latter means no sample can avoid all marked slots.

Recovering the two original probabilities. For completeness, no marked slot is sampled exactly when $W \subseteq C$. Thus,

$$P_{\text{cover}}(c \mid m, N) = \Pr(W \subseteq C) = \frac{\binom{c}{m}}{\binom{N}{m}} = 1 - P_{\text{det}}(N, N - c, m). \tag{10}$$

This event tests coverage of true facts. A purely omitted slot belongs to $\mathcal{U} \setminus C$, but so does a slot at which an incorrect value replaces the true fact. For an omission-only analysis one would instead mark the set of unexpressed slots; that is a different event and cannot silently replace $\mathcal{U} \setminus C$ in the original coverage formula.

For correctness, no erroneous assertion is exposed exactly when $W \cap H = \varnothing$. Hence,

$$P_{\text{miss}}(e \mid m, N) = \Pr(W \cap H = \varnothing) = \frac{\binom{N-e}{m}}{\binom{N}{m}} = 1 - P_{\text{det}}(N, e, m). \tag{11}$$

Both calculations sample reference slots, not candidate assertions. The two marked sets can overlap, and both checks use the same W; no independence is assumed and the marginal probabilities are not multiplied into a joint reward. If a candidate makes both a true and a false assertion at a slot, that slot may belong to both C and H: the correct fact is covered, while correctness can still fail. The model counts slots with defects, not all individual erroneous statements.

Sensitivity and graded supervision. By linearity of expectation, $\mathbb{E}[K_D] = md/N$, since each marked slot has inclusion probability m/N. For fixed N and d, increasing m makes detection nondecreasing: couple the samples as prefixes of a uniform random permutation of $\mathcal{U}$. Once a prefix hits a marked slot, every longer prefix also does. For fixed N and m, increasing d is similarly nondecreasing by coupling nested marked sets.

Weak reference coverage and strong defect detection. For $1 \le d, m \le N$ and $m \le N - d$, sampling without replacement gives

$$\Pr(K_D = 0) = \prod_{j=0}^{m-1} \left(1 - \frac{d}{N-j}\right) \le \left(1 - \frac{d}{N}\right)^m.$$

If $m > N - d$, the missed-detection probability is zero and the bound still holds. For $m \ge 1$, every nonempty defect set has a positive probability of being probed, so $P_{\text{det}} = 0$ if and only if $d = 0$. The ideal defect-free criteria are therefore $c = N$ and $e = 0$, independently of reference size; this does not establish the optimum of the implemented weighted reward. For a defect fraction $\rho = d/N \in (0, 1)$ and target miss probability $\delta \in (0, 1)$, any feasible $m \ge \lceil \log(1/\delta)/[-\log(1-\rho)] \rceil$ ensures $P_{\text{det}} \ge 1 - \delta$. At fixed ρ, this sufficient probe count depends on confidence and defect density rather than total scene size. For example, $N = 1000$, $d = 100$, and $m = 30$ give 95.97% detection while probing only 3% of the slots. With $N = 20$ and $d = 4$, $m = 5$ and 10 give 71.8% and 95.7%. These are ideal-model calculations for either defect set. They explain how incomplete references can expose defects reliably; rare defects require more probes, and reliable exposure does not itself guarantee judge accuracy or policy improvement.

Marginal responses to candidate quality. These detection monotonicities do not imply uniformly larger marginal responses to a change in candidate quality. The original finite differences, for $1 \le m \le N$, are

$$P_{\text{cover}}(c+1) - P_{\text{cover}}(c) = \frac{\binom{c}{m-1}}{\binom{N}{m}}, \quad 0 \le c < N, \tag{12}$$

$$P_{\text{miss}}(e) - P_{\text{miss}}(e+1) = \frac{\binom{N-e-1}{m-1}}{\binom{N}{m}}, \quad 0 \le e < N. \tag{13}$$

They follow from Pascal's identity. The first is positive when $c \ge m-1$; the second when $e \le N-m$. For $N = 10$ and c increasing from 1 to 2, the coverage-probability increments for $m = 1, 2, 4$ are 0.1, $1/45$, and 0. In particular, $c < m$ gives zero coverage probability regardless of partial improvements that remain below m. Counts of exposed defects retain information that such binary events discard. VCap uses a qualitative rubric to map the number and significance of confirmed omissions and errors to graded scores; the rubric does not prescribe an exact numerical penalty per fact. Neither the event probabilities nor their finite differences establish the policy's optimization gradients, convergence, or absence of a length-dependent tradeoff.

Unequal reference coverage. For nonuniform W, define $q_i = \Pr(i \in W)$. Writing the intersection count as a sum of indicators gives

$$\mathbb{E}[K_D] = \sum_{i \in D} q_i, \tag{14}$$

and, for the original correct and erroneous slot sets,

$$\mathbb{E}|W \cap C| = \sum_{i \in C} q_i, \qquad \mathbb{E}|W \cap H| = \sum_{i \in H} q_i. \tag{15}$$

These identities require no independence across slot inclusions. A slot with $q_i = 0$ receives no direct supervision through this channel. The marginal inclusion probabilities alone generally do not determine $\Pr(K_D \ge 1)$; the hypergeometric formula requires uniform sampling of size-m subsets. A fixed natural reference pool need not satisfy this assumption. Reference noise, imperfect visual judgments, and ordinal score aggregation introduce additional effects. The actual judge may inspect facts outside the reference, but the reliability of that additional channel is an empirical property.

A.3 Related Work

A.3.1 Visual Captioning

Visual captioning has evolved from short single-sentence descriptions on MS-COCO (Chen et al., 2015) and Flickr30K (Young et al., 2014) toward dense, fact-rich descriptions of images and videos.

Recent progress has largely relied on scaling caption supervision through large image-text corpora and synthetic recaptioning pipelines. Early web-scale datasets such as LAION (Schuhmann et al., 2022), CC12M (Changpinyo et al., 2021), and YFCC100M (Thomee et al., 2016) enabled large-scale vision-language pretraining, but their alt-text annotations are typically short, noisy, and fact-incomplete.

To improve supervision quality, many works refine or replace noisy captions with stronger synthetic descriptions. BLIP (Li et al., 2022) uses caption generation and filtering to bootstrap training data; BLIP-2 (Li et al., 2023a) connects frozen image encoders and language models through a two-stage pretrained Querying Transformer. LaCLIP (Fan et al., 2023) rewrites captions with LLM assistance to improve semantic alignment, while CapsFusion (Yu et al., 2024a) consolidates raw alt-text and synthetic captions into denser supervision. A parallel direction distills supervision from frontier proprietary models or human annotators. ShareGPT4V (Chen et al., 2023), ShareGPT4Video (Chen et al., 2024c), and ALLaVA (Chen et al., 2024a) construct large-scale dense caption datasets using GPT-4V-style systems, while DCI (Urbanek et al., 2024), ImageInWords (Garg et al., 2024), PixMo-Cap (Deitke et al., 2024), DenseFusion-1M (Li et al., 2024), and PixelProse (Singla et al., 2024) further improve caption density through human annotation or curated recaptioning. Similar trends extend to video captioning, including Panda-70M (Chen et al., 2024d), MiraData (Ju et al., 2024), Vript (Yang et al., 2024), OpenVid-1M (Nan et al., 2024), and Tarsier2-Recap (Yuan et al., 2025).

Another line of work improves caption quality through inference-time composition rather than re-training. These approaches combine object detectors, region-level captioners, or external LLM agents to improve coverage and factual grounding (Sun et al., 2024a; Lee et al., 2025). While such systems often improve factual coverage, they introduce substantial inference complexity and still ultimately depend on the quality of their constituent captioning modules.

Recent evaluation research has increasingly shifted from holistic similarity metrics toward fact-level assessment. FaithScore (Jing et al., 2024) grounds atomic claims against images, CompreCap (Lu et al., 2024) models captions as structured scene graphs, and CapMAS (Lee et al., 2025) evaluates factual correctness and coverage jointly. On video captioning, DREAM-1K (Wang et al., 2024b), VDC (Chai et al., 2025), CAPability (Liu et al., 2025), and ARGUS (Rawal et al., 2025) similarly decompose evaluation into precision- and recall-oriented dimensions. Standard hallucination benchmarks (Li et al., 2023b; Wang et al., 2023; Guan et al., 2024; Sun et al., 2024b) also remain widely used. Despite this transition toward fact-level evaluation, most caption training objectives still optimize imitation or holistic preference signals, leaving the gap between factual evaluation and factual supervision unresolved.

A.3.2 Reinforcement Learning for Vision-Language Models

Reinforcement learning has become a dominant paradigm for post-training large language models. RLHF (Ouyang et al., 2022) demonstrated that preference optimization can substantially improve model behavior beyond supervised fine-tuning, while RL with verifiable rewards (RLVR) (Shao et al., 2024; DeepSeek-AI et al., 2025) extended this paradigm to domains such as mathematics and code generation. GRPO (Shao et al., 2024) has since emerged as a widely adopted optimizer for on-policy RL training.

These ideas have increasingly transferred to vision-language models (Lu et al., 2026). Task-specific verification includes CapRL's VQA-based caption rewards (Xing et al., 2025), AC-RL's supervision through downstream task success and clarification needs (Gkountouras & Titov, 2025), and RLVR for visual reasoning, grounding, and spatio-temporal perception (Shen et al., 2025; Li et al., 2025). Preference optimization reduces hallucination and improves alignment (Yu et al., 2024b; 2025; Xiao et al., 2025; Wang et al., 2024a; Xie et al., 2024). ViCrit (Wang et al., 2025b) trains a visual critic to locate injected caption errors as a proxy task for perception. Caption-generation methods instead use scene-graph correction rewards in SC-Captioner (Zhang et al., 2025a), sample-specific rubric scoring in RubiCap (Huang et al., 2026), and caption-set equivalence in OwlCap (Zhong et al., 2025). Related video-captioning approaches include VideoCap-R1 (Meng et al., 2025), video-SALMONN2 (Tang et al., 2025), VidBridge-R1 (Chen et al., 2025), and VDC-Agent (Wang et al., 2025a). FeedQuill and DCScore (Ye et al., 2025) further decompose responses into atomic units and optimize factual precision with length-aware preference signals.

VCap shares the goal of factual supervision with these approaches. Its focus is the complementary use of reference text and visual evidence within one reward interface, evaluated for image and video captioning. The fact-slot analysis describes the reference-guided channel under explicit assumptions and does not establish an unconditional advantage over every existing reward.

A.4 Reference Roles Across Reward Families

Reference matching and preference learning. Similarity-based rewards compare a candidate with one or more references and can favor reference overlap over valid additional details. However, the optimum depends on the metric, reference set, and candidate space. Preference optimization is a training objective whose supervision depends on the preference construction; it is not generally equivalent to copying a reference. We therefore do not assign a universal reference-quality ceiling to SFT, DPO, or every similarity metric.

Question-answering rewards. VQA-based rewards inspect whether a caption supports answers to a question pool. Their coverage depends on how questions are chosen, and facts outside the pool may receive little supervision. A reference-guided reward has an analogous coverage issue: reference facts are also selective. VCap additionally asks a judge to use visual evidence in assessing the candidate, but the success of that assessment depends on the judge.

Model-based factual judgments. A multimodal judge can already use images, references, and decomposed criteria. VCap belongs to this broad family. The witness–adjudicator formulation specifies the intended information roles and motivates a shared image/video scoring interface. Its empirical value is examined through input and score ablations. The idealized analysis neither proves that all earlier judges lack factual localization nor substitutes for a matched comparison with a similarly configured judge.

A.5 Video-Captioning Reward-System Ablation

We mirror the image-captioning ablation of Section 3.4 on the video side: starting from VCap (E1) we retrain under five lesions, one removing the reference caption (the reward model sees the video only), three zeroing one of $w_{\text{corr}}/w_{\text{comp}}/w_{\text{txt}}$ before aggregation, and a video-specific one zeroing w_{local} in Equation (6) so that the per-segment local pass no longer contributes to the sentence-level reward. Table 5 reports VCapsBench and VDC scores for the reported ablation configurations. Section 3.4 discusses their contributions and the remaining error-control limitation.

A.6 Instructions for VCap Reward Modeling

This subsection collects the four instructions with which the frozen reward model of Section 2.2 is queried: two for image captioning (Figures 4 and 5) and two for video captioning (Figures 6 and 7). Each instruction asks the same reward model to (i) compare the Generated Caption against the visual signal using the Reference Caption only as auxiliary information, (ii) emit a concise analysis listing confirmed problems, and (iii) return three integer scores in $\{0, 1, \ldots, 10\}$ inside a single JSON object. To save space, the second instruction in each pair only spells out the parts that differ from the first.

A.6.1 Image Captioning Instructions

Instruction 1: Image Caption Reward (with reference)

You are an expert in image captioning. I will provide a Reference Caption and a Generated Caption for the Input Image above. Use the Input Image as ground truth and the Reference Caption as auxiliary context to evaluate the Correctness, Completeness, and Text Quality of the Generated Caption.
Principles.

- The Image is the source of truth. If the Reference Caption conflicts with the Image, prioritize the Image; the Reference Caption is auxiliary only.
- Pay more attention to salient content (main subjects, key attributes such as count/color/type, actions, crucial spatial relations) than to background minutiae.

Operation steps.

1. Analyze the Generated Caption on three dimensions:
   - **Correctness**: does it accurately represent the image, free of objects not present, inaccuracies, or contradictions? Fewer mistakes is better.
   - **Completeness**: does it cover all objects in detail with no omissions of details or objects? Fewer omissions is better.
   - **Text Quality**: is it logically fluent, coherent, concise (no repetitions), and aesthetically pleasing? Penalize any meta-text unrelated to the image content, e.g. self-evaluations such as "all elements have been described" or "no detail or object is omitted."
2. Provide an integer score in $\{0, 1, \ldots, 10\}$ for each dimension.

Input.
Reference Caption: `ref_answer`
Generated Caption: `gen_solution`
Output format (strict). Return exactly one JSON object: `{"Analysis": ⟨your analysis⟩, "Correctness": score1, "Completeness": score2, "Text Quality": score3}`

Figure 4: Reward-model instruction for image captioning when a reference caption is available.

Reward computation flow. Each instruction is instantiated with the policy caption and reference, then submitted to the frozen judge with the image, timestamp-tagged full-video frames, or the relevant segment's frames in one multimodal forward pass. The no-reference ablation omits the reference slot. We parse the three integer scores from its JSON response and map them to correctness, completeness, and the auxiliary dimension in Equation (1). Text Quality is auxiliary for images and segments; Reasonability is auxiliary for the global video pass. Video combines global and local rewards by Equation (6) before the GRPO update. Local scoring reuses the full-video reference and compares the corresponding facts in the generated description.

Instruction 2: Image Caption Reward (no reference, ablation)

You are an expert in image captioning. I will provide a Generated Caption for the Input Image above. Use the Input Image as the sole source of truth to evaluate the Correctness, Completeness, and Text Quality of the Generated Caption.
Dimension definitions and operation steps are identical to Instruction 1, with two differences: (i) no Reference Caption is provided, so the cross-check between the two captions is dropped and the analysis grounds directly in the image; (ii) the Text Quality criterion additionally penalizes courtesy text, modification notes, personal feedback for suggestions, and self-evaluations such as "all the elements of the image have been clearly described" or "no detail or object is omitted/overlooked".
Input.
Generated Caption: `gen_solution`
Output format. Same JSON object as Instruction 1.

Figure 5: Reward-model instruction for image captioning without a reference caption, used in the no-reference ablation of Section 3.4.

A.6.2 Video Captioning Instructions

Instruction 3: Video Caption Reward (global)

You are an expert video-description quality evaluator. The frames above are extracted from a video, each tagged with a timestamp (in seconds). You are given a Reference Description as auxiliary information and a Generated Description to be evaluated. Use the video frames as the **only** ground truth and rate the Generated Description on three dimensions: Reasonability, Correctness, and Completeness.
Principles.

- Be evidence-based: only contents visible in the frames count as ground truth. If the Reference Description and the Generated Description disagree, judge by the frames.
- Pay main attention to salient content (main subjects, key attributes, key actions/events over time, important interactions/spatial relations); more details are preferred.

Definitions.

- **Reasonability (segmentation)**: are the segment splits appropriate, with each segment having a coherent and relatively independent topic, and with boundaries aligned to visible content changes?
- **Correctness**: does the Generated Description avoid hallucinations, inaccuracies, contradictions, and timestamp mismatches in each segment's description and in the global summary? Are the time ranges of important events labelled accurately?
- **Completeness**: does it cover all major visible entities, attributes, fine-grained details, and events, both per segment and in the global summary?

Operation steps.

1. Read the Generated Description; inspect frames in timestamp order; compare against the Reference Description and decide by frames on disagreements.
2. List confirmed hallucinations, inaccuracies, omissions, and timestamp mismatches per dimension.
3. Provide an integer score in $\{0, 1, \ldots, 10\}$ for each dimension.

Input.
Reference Description: `REF_DESC`
Generated Description: `GEN_DESC`
Output format (strict). Return exactly one JSON object: `{"Analysis": ⟨your analysis⟩, "Reasonability": score1, "Correctness": score2, "Completeness": score3}`

Figure 6: Reward-model instruction for the global pass over a full video, scoring Reasonability, Correctness, and Completeness.

Instruction 4: Video Caption Reward (per-segment local)

You are an expert video-description quality evaluator. The frames above belong to a single video *segment*, each tagged with a timestamp. You are given a Reference Description for the *whole* video as auxiliary information and a Generated Description. Evaluate its content corresponding to this segment, using the segment frames as the only ground truth, and rate Correctness, Completeness, and Text Quality.
Principles, evidence-based instruction, and operation steps follow Instruction 3, with two differences: (i) the Reference Description spans the whole video and contains content outside the segment; the model is instructed to ignore such out-of-range parts; (ii) the Reasonability dimension is dropped (a single segment carries no segmentation to evaluate) and replaced with **Text Quality**, which checks fluency, coherence, conciseness (no repetitions), aesthetic quality, and absence of text unrelated to the video content.
Input.
Reference Description (whole video): `REF_DESC`
Generated Description: `GEN_DESC`
Output format (strict). Return exactly one JSON object: `{"Analysis": ⟨your analysis⟩, "Correctness": score1, "Completeness": score2, "Text Quality": score3}`

Figure 7: Local reward instruction for a segment defined by the captioner, using the full-video reference and evaluating only the corresponding facts.

A.7 Case Study: Reference, VCap (e1), and VCap (e2)

We present the complete captions from the original backbone, VCap (e1), and VCap (e2) on one COCONut training image. Both e1 and e2 start from the original backbone; e2 uses references generated by e1. This example illustrates changes in description, while held-out benchmarks and human assessment provide quantitative evidence.

Reference caption: original backbone

This is a street-level photograph capturing a scene of urban springtime, dominated by a large, flowering tree in the foreground. The tree, likely a cherry or plum variety, has a thick, dark-barked trunk and a sprawling canopy of delicate, pinkish-white blossoms intermingled with reddish-brown leaves, suggesting

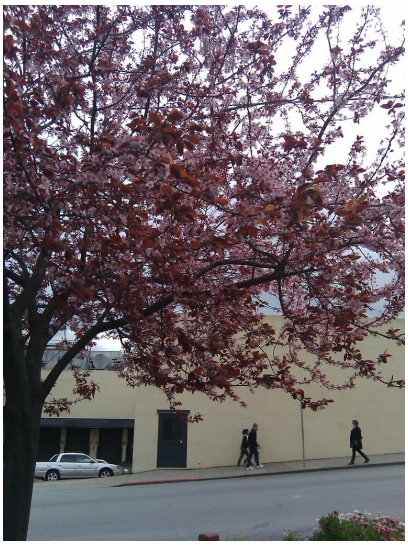


Figure 8: Training-set example used for the caption comparison: a flowering tree, pedestrians, and a building along a street.

Reference caption: original backbone (continued)

it might be a cultivar like the “Autumnalis” or “Kanzan” cherry. The branches arch gracefully over the scene, partially obscuring the sky and casting dappled shadows. The blossoms appear dense and soft, creating a beautiful contrast against the darker branches and the overcast, pale sky in the background.
Below the tree, a paved street stretches across the lower part of the image. A curb painted red runs along the edge of the sidewalk. A silver pickup truck is parked on the left side of the street, angled slightly toward the viewer. Behind the truck and the tree, a long, single-story building with a pale yellow or beige stucco exterior forms the backdrop. A dark blue or black door with a small window is visible on the building’s facade, and above it, a row of windows or openings can be seen.
On the sidewalk, three pedestrians are walking from left to right. The person on the far right is walking alone, wearing a dark coat and dark pants, and appears to be carrying a bag. To their left, two other individuals are walking side by side; one is wearing a dark jacket and jeans, while the other, slightly ahead, is wearing a dark top and lighter-colored pants. Their movement adds a sense of life and motion to the scene.
In the bottom right corner, a small patch of greenery with pink and white flowers can be seen, possibly from a flowerbed or planter, adding a splash of color to the edge of the frame. The overall lighting is soft and diffused, typical of an overcast day, which enhances the pastel tones of the blossoms and the muted colors of the surroundings. The image evokes a quiet, contemplative moment in a city neighborhood during the spring season.

Figure 9: Complete caption from the original backbone, used as the initial reference.

VCap (e1): complete caption

This is an outdoor, street-level photograph capturing a tranquil, everyday urban or suburban scene framed by the prominent natural beauty of a flowering tree.
Overall Composition and Mood. The image is framed with a strong sense of depth and layered elements. The overarching mood is one of quiet transition—the arrival of spring juxtaposed with the routine of daily life. The lighting is soft and diffused, characteristic of an overcast day, which casts even tones across the scene and enhances the colors and textures without harsh shadows. The composition is anchored by the massive, sprawling form of the tree in the foreground, which acts as a natural frame and visual weight, guiding the viewer’s eye toward the background activity.
Foreground—the dominant tree. The tree occupies the entire left and central portion of the frame, with its trunk visible on the far left edge and its branches and blossoms cascading across the upper and middle portions of the image. The visible bark on the left side appears dark and textured, grounding the composition. The tree is in full bloom, covered in dense clusters of small, delicate blossoms in a soft, pastel pink or light lavender hue, creating a gentle, ethereal wash. Intermingled with the flowers are

VCap (e1): complete caption (continued)

young, emerging leaves that are not green but a striking, deep burgundy, maroon, or reddish-brown—a key visual feature that creates a rich, warm contrast with the cool pink blossoms and the pale sky. The dark, intricate network of branches and twigs forms a complex, organic lacework over the background, partially obscuring and framing the elements behind it.
Foreground—street and sidewalk. A wide asphalt road stretches horizontally across the lower third of the image; a narrow, reddish-pink painted curb line runs parallel to the sidewalk along the edge of the road. A concrete sidewalk is visible along the base of the beige building, where the pedestrians are walking.
Midground—pedestrians. Three people walk along the sidewalk in the midground, moving from right to left. On the far right, a single person walks alone in dark, likely black, outerwear (a long coat or jacket) and dark pants; the posture is upright and purposeful. Closer to the left, two people walk side by side and appear to be companions: the person on the left wears a dark top and dark pants, while the person on the right (closer to the building) wears a dark jacket or top with a noticeable white or light-coloured horizontal band or trim across the chest/belly area. Both are dressed for cooler weather, and the spacing of the trio suggests casual, unhurried movement.
Midground—vehicle. Parked along the left side of the street, just before the sidewalk begins, is a silver or light gray pickup truck. Its position is partially obscured by the lower branches of the large tree and the building's structure, integrating it into the layered composition.
Background—building. A large, single-story building forms the dominant background element. Its walls are painted a uniform, light beige or off-white colour. A distinct, dark blue or black rectangular door with multiple panels is set into the wall, positioned centrally in the lower midground. To the left of the door, a dark, recessed area appears to be a garage entrance or covered walkway; visible within this recess are vertical structural elements, and above it, several HVAC units or circular mechanical vents are mounted on the building's exterior wall. The building's design is functional and unadorned, serving as a neutral canvas that allows the tree and the people to stand out.
Background—sky. The sky, visible only through the gaps in the tree's canopy, is a uniform, pale, overcast white or very light gray, providing a soft, non-distracting backdrop that allows the colours of the blossoms and leaves to be prominent.
Additional details. The dominant colours are the muted beige of the building, the soft pink of the blossoms, the rich burgundy of the leaves, the dark tones of the trunk, branches and pedestrian clothing, and the gray of the road and overcast sky—together a harmonious, subdued palette with accents of warmth from the foliage. A thin, gray utility pole is visible on the far right side of the building, just to the right of the central group of pedestrians. In the bottom right corner, a small portion of a bush or flower bed with pink and white flowers and green foliage is visible, anchoring the bottom right and providing a touch of repetition with the tree in the upper frame.

Figure 10: Complete VCap (e1) caption, trained from the original backbone using its references.

VCap (e2): complete caption

This image presents a layered urban street scene viewed through the dominant presence of a flowering tree, creating a composition that blends natural beauty with everyday human and architectural elements.
Overall impression and composition. The photograph captures a candid, observational moment on a street corner or sidewalk. The framing is deliberately angled and partially obscured by the foreground subject, lending an intimate and slightly voyeuristic perspective. The overall mood is quiet and contemplative, underscored by the soft, diffused lighting and the gentle movement of the pedestrians. There is a harmonious yet contrasting interplay between the organic, colourful vibrancy of the tree and the muted, structured geometry of the built environment.
Foreground—the dominant tree. The tree occupies the left and central portions of the frame, its trunk anchoring the lower-left corner and its sprawling branches arching diagonally across the upper half and right side; its sheer size and proximity make it the undeniable primary subject. The visible trunk is dark, textured, and substantial, providing a grounding element against the lighter background, transitioning into a complex network of branches that extend outward and upward in intricate silhouettes against the sky. The tree is in a state of unique bloom: leaves and blossoms exhibit a rich, deep reddish-brown or maroon hue rather than the typical green or pink, creating a striking, almost autumnal colour statement; the dense canopy is dotted with soft, pale pink or whitish blossom clusters nestled among the darker leaves, and the interplay of light and shadow within the foliage creates depth and texture. The dark, sinuous branches form strong leading lines that draw the eye inward and upward, framing the scene behind them and partially veiling the elements beyond.
Midground—pedestrians and infrastructure. To the right of the tree's lower branches, three pedestrians walk along the sidewalk, captured mid-stride; two individuals walk closely together, slightly ahead of the third, all dressed in dark, casual or business-casual attire (coats, pants) that contrasts with the warmer tree

VCap (e2): complete caption (continued)

tones and blends into the neutral background wall. Their small scale relative to the tree and the building emphasizes the dominance of the natural element. Further to the right, a single pedestrian walks alone, slightly further down the sidewalk—a solitary figure that adds to the narrative of quiet urban movement.
Midground—architecture. Behind the pedestrians and the lower branches, a large, plain building wall forms the backdrop, painted a uniform light beige or pale yellow that provides a neutral canvas. A dark rectangular door (possibly blue or black) is visible on the wall, offering a point of contrast in colour and form; to the left, beneath the overhanging branches, a darker recessed area or awning structure can be seen at ground level, and higher up near the roofline behind the upper canopy, faint details that look like HVAC units or vents are visible, grounding the building in its functional, real-world context.
Midground—vehicle. Parked along the curb on the left side of the street, partially sheltered by the tree's lower branches and the building's recessed area, is a silver or light-coloured pickup truck whose presence firmly situates the scene in a functional, accessible urban or suburban setting.
Background and atmosphere. The sky visible through the gaps in the canopy is a flat, overcast white or very pale gray; this diffuse lighting eliminates harsh shadows and contributes to the soft, serene, slightly melancholic atmosphere, enhancing the ethereal quality of the blossoms and the muted tones of the environment. The asphalt street occupies the bottom portion of the frame, its gray surface providing a stable horizontal base, and a distinct red or reddish-brown painted curb separates the sidewalk from the street, adding a small but important splash of saturated colour and defining the boundary of the human realm. At the very bottom right corner, a hint of additional landscaping is visible—low-lying greenery and a cluster of small, likely pink or white blossoms—subtly echoing the main tree's floral theme.
Colour palette and visual dynamics. The scene is defined by a sophisticated interplay of earthy and muted tones: deep red/brown in the leaves and blossoms, soft pink/white in the delicate flowers, beige/yellow in the wall, dark blues/blacks in the door, pedestrian clothing and tree branches, gray/white in the sky and street, and red/orange on the curb line. There is strong contrast between the dark, structural lines of the branches and the lighter background and between the vibrant foliage and the neutral architecture, but the overall palette is harmonious due to its earthiness and shared tones. The image layers multiple textures—rough bark, soft petals and leaves, smooth wall, asphalt grain, fabric—and combines them with overlapping elements (branches over building, pedestrians on sidewalk) to create profound depth and a tangible sense of space.
Narrative and mood. The quiet passage of people against the backdrop of a uniquely blossoming tree suggests themes of fleeting moments, seasonal change, and the coexistence of nature within the constructed urban world. The overcast sky lends a pensive, calm quality, inviting the viewer to pause and appreciate the subtle beauty found in an ordinary street corner; the tree acts as both a shelter and a frame, turning a mundane scene into something quietly poetic and visually rich, balancing the ephemeral (blossoms, passing people) with the permanent (building, street).

Figure 11: Complete VCap (e2) caption, trained from the original backbone using e1-generated references.

Interpretation. The outputs add architectural and clothing details but also disagree about walking direction: the reference says left-to-right, whereas e1 says right-to-left. The e2 caption first describes three pedestrians and later mentions a solitary person, leaving their grouping ambiguous. These differences illustrate changes in coverage and specificity, rather than verified gains for every added detail. The complete captions also contain subjective descriptions of mood and composition, excluded from factual annotation. Quantitative evidence comes from the benchmarks and human assessment.

A.8 Human Annotation Protocol

Each unit contains one image and five captions: VCap (e2), Seed 2.0 Pro, Gemini 3.1 Pro, GPT-5.4, and Qwen3.5-397B. Qwen3-VL-235B-Instruct proposes missing and inconsistent facts for human verification, followed by three per-image rankings.

Step 1: proposition vs. image. Without consulting the caption, annotators judge the proposition as `1. consistent`, `2. inconsistent`, or `3. undecidable`. The last label covers unresolved visual evidence such as occlusion, low resolution, or ambiguous viewing angle; clear evidence requires a definite verdict. For "the image shows A, but the caption says B," this step judges only the image-side assertion A.

Step 2: proposition vs. caption. Only Step-1 `1. consistent` propositions proceed; otherwise this field is blank. Annotators compare the image with the relevant caption spans and choose `1. not holding`, `2. holding`, or `3. ambiguous`. A missing proposition holds when the content is absent from the entire caption, which must be read before confirming an omission. Synonyms and hypernyms count as entity mentions, but attributes such as color and count must match at their own granularity. An inconsistent proposition holds when the alleged erroneous wording is present, including in a self-contradictory caption. It does not hold if that wording is absent or misrepresents a more accurate description. Vague wording that prevents a decision receives `3. ambiguous`.

Step 3: per-image rankings. After labeling all five captions, annotators independently rank accuracy, completeness, and granularity. Accuracy and completeness use confirmed inconsistent and missing propositions, respectively; granularity uses the detail in the caption. Each ranking is a strict permutation of $\{1, \ldots, 5\}$, with 1 best and ties resolved by the annotator's overall judgment. Rankings are produced once and are not retroactively edited.

Scope and consistency. Only objective visual facts are in scope: objects, attributes, counts, positions, text, colors, actions, and spatial relations. Mood, aesthetic judgments, narrative interpretation, and photographic intent receive Step-1 `3. undecidable` with the note "subjective." Each proposition is judged independently and is not reused across captions. Labels retain the exact `1./2./3.` prefixes for downstream parsing.

Quick reference. Table 9 summarizes the most common proposition–image–caption configurations and the (Step 1, Step 2) labels they map to, mirroring the cheat sheet given to annotators.

Table 9: Quick reference for the two-step proposition labels used in the human annotation protocol.

Configuration	Step 1	Step 2
Image really contains the asserted content; caption indeed omits it	`1. consistent`	`2. holding`
Image really contains the asserted content; caption already covers it via a synonym/hypernym	`1. consistent`	`1. not holding`
Image really contains the asserted content; caption is too vague to decide	`1. consistent`	`3. ambiguous`
Image does not contain the asserted content	`2. inconsistent`	—
Object is occluded / blurred / unresolvable in the image	`3. undecidable`	—
Proposition: "image shows A"; image is indeed A; caption says B	`1. consistent`	`2. holding`
Proposition: "image shows A"; image is indeed A; caption also says A (proposition misjudged)	`1. consistent`	`1. not holding`
Proposition: "image shows A"; image actually shows C ($\neq A, \neq B$)	`2. inconsistent`	—