Re^{3}Cap: Retrieval-Guided Refinement for Image Captioning Enhancement via Reinforcement Learning

Haonan Jia[1]*, Shichao Dong[1]*, Zenghui Sun[1], Jiawen Zheng[2], Ziqi Miao[3], Gege Shi[1], Qiuyu Zhao[1], Jinsong Lan[1], Xiaoyong Zhu[1], Bo Zheng[1]†

[1]Taobao & Tmall Group of Alibaba

[2]The Hong Kong University of Science and Technology (Guangzhou)

[3]Shanghai Artificial Intelligence Laboratory

arXiv:2608.21305v1 [cs.CV] 21 Aug 2026

Abstract

Reinforcement Learning (RL) has demonstrated significant gains in image captioning, yet it is still limited in encouraging Large Vision-Language Models (LVLMs) to explore novel reasoning strategies. This limitation leads to a performance gap between RL and Supervised Fine-Tuning (SFT). In this paper, we argue that multi-modal retrieval can serve as an effective reasoning signal for caption refinement. Based on this insight, we present the Retrieval-Guided Refinement for Image Captioning (Re^{3}Cap), a retrieval-guided reasoning strategy that enhances image captioning without requiring additional annotations. Instantiated by Caption Refinement Suggester (CRS) and Caption Quality Assessor (CQA), this strategy identifies hallucinations and omissions in image captions, leading to more accurate and detailed descriptions. Extensive experiments demonstrate the superiority of our method in image captioning, even compared with Supervised Fine-Tuning. Especially, Re^{3}Cap outperforms GRPO with an average improvement of 8.64% in relation reasoning on the COCO-LN500 benchmark. The code will be released when the paper is accepted.

1 Introduction

Image captioning (Karpathy and Fei-Fei, 2015; Huang et al., 2019; Liu et al., 2017b) is a fundamental task in computer vision and plays an essential role in various applications, such as text-image retrieval (Duan et al., 2025; Chen et al., 2024b), text-to-image generation (Betker et al., 2023; Zheng et al., 2024), and visual question answering (Cheng et al., 2025; Hu et al., 2024; Miao et al., 2026). Recently, the development of Large Vision-Language Models (LVLMs) (Dai et al., 2023; Liu et al., 2023; Dong et al., 2024; Bai et al., 2023; Ye et al., 2023) has demonstrated notable success in multi-modal understanding and yielded significant performance gains in image captioning. Nevertheless, image captions generated by existing methods (Cornia et al., 2020; Huang et al., 2019; Liu et al., 2017a,b; Feng et al., 2019; Bahng et al., 2025; Tewel et al., 2022; Xu et al., 2023) are prone to hallucinations and often fail to capture fine-grained visual details. Consequently, it remains challenging to generate detailed and accurate image captions.

Previous studies have primarily leveraged reinforcement learning (RL) to post-train Large Vision-Language Models (LVLMs) to enhance their image captioning capabilities. For instance, CLIP-based methods (Cho et al., 2022; Yu et al., 2023; Dzabraev et al., 2024) assess the correlation score between images and LVLM-generated captions based on Vision Language Models (VLMs). By using this score as the reward signal, these methods force LVLMs to generate more detailed image captions. Unfortunately, due to the constrained compositional reasoning capabilities of VLMs (Wang et al., 2024a), these methods remain highly susceptible to reward hacking. Accordingly, SC-Captioner (Zhang et al., 2025) annotates keywords for each image and evaluates the quality of image captions by checking whether the caption explicitly contains these words. However, RL-based approaches still lag behind Supervised Fine-Tuning methods (Luo et al., 2024; Yang et al., 2025b; Li et al., 2024b; Chen et al., 2024a).

Recent studies (Yue et al., 2025) reveal that reinforcement learning merely selects the highest-reward caption from candidates that are pre-generated by LVLMs. During training, LVLMs exhibit limited exploration of novel reasoning strategies, failing to produce diverse and previously unexplored caption candidates. In this case, the captioning performance of RL-optimized LVLMs remains bounded by the intrinsic reasoning capabilities of the base model. Consequently, the crucial challenge in image captioning lies in exploring novel

*Equal contribution.

†Corresponding author.

reasoning strategies that empower models to generate unexplored candidate captions for RL.

In this paper, we argue that multi-modal retrieval can serve as an effective reasoning signal for caption refinement. Intuitively, **visually similar images often share overlapping semantic content**. In this way, by using the source image as a query, we can infer its semantic content from descriptions of the retrieved similar images. Moreover, **semantically similar queries tend to yield consistent retrieval results**. Consequently, a detailed and accurate image caption should induce retrieval results similar to those obtained from the source image. Any discrepancy between image-based and caption-based retrieval descriptions signals the misalignment between the image and the LVLMs-generated caption, indicating hallucinations.

Building on these observations, we introduce Re^{3}Cap, a retrieval-guided reasoning strategy that enhances image captioning through two components: the Caption Refinement Suggester (CRS) and the Caption Quality Assessor (CQA). Specifically, CRS identifies critical elements to preserve in the image caption by verifying descriptions that consistently overlap across retrieved similar images. Furthermore, CQA analyzes discrepancies between image-based and caption-based retrieval descriptions to indicate hallucinations and omissions in the generated caption. Through this process, we determine which elements in LVLM-generated captions should be retained, which hallucinated content needs to be removed, and which visual details from the source image have been omitted. During RL training, such reasoning results will guide LVLMs to refine their captions without requiring additional annotations. By injecting this reasoning strategy, LVLMs generate diverse, previously unexplored caption candidates, thereby significantly enhancing their image captioning capability. Extensive experiments demonstrate the effectiveness of our proposed method across multiple LVLM architectures under diverse reward functions. Our contributions can be summarized as follows:

- We present a novel retrieval-based reasoning strategy to indicate hallucinations and omissions in captions without requiring additional annotations.
- We propose Re^{3}Cap, which guides LVLMs to generate previously unexplored caption candidates, thereby enhancing model performance in image captioning.
- Extensive experiments demonstrate that our method improved the performance on various LVLMs, outperforming state-of-the-art methods by a large margin, even compared with Supervised Fine-Tuning.

2 Related Work

Image captioning is a fundamental task in computer vision, serving as a key bridge between the visual and linguistic modalities. Recent works can be broadly categorized into two lines: supervised fine-tuning (SFT) and reinforcement learning (RL).

2.1 Supervised Fine-Tuning

Previous approaches typically adopt an encoder–decoder paradigm, where an encoder extracts visual representations from the image, and a decoder autoregressively generates the caption (Cornia et al., 2020; Huang et al., 2019; Liu et al., 2017a,b; Vinyals et al., 2015; Wang et al., 2022; Mokady et al., 2021; Luo et al., 2023). Building upon the encoder–decoder paradigm, several works further incorporate retrieval-augmented generation (RAG), where the captioner is conditioned not only on the image but also on relevant texts retrieved from external corpora (Ramos et al., 2023; Li et al., 2024a; Kim et al., 2025). Complementary approaches substitute human annotations with synthetic data for supervision (Luo et al., 2024; Yang et al., 2025b; Li et al., 2024b; Chen et al., 2024a). In addition, some approaches perform self-supervised training by leveraging the shared multimodal embedding space of vision–language models (Fei et al., 2023; Tam et al., 2023; Lee et al., 2025). Controllability has also been explored by fine-tuning captioning models to obey user-specified control signals (Kornblith et al., 2023; Saito et al., 2025). Despite substantial gains in caption accuracy and detail, these methods still heavily depend on large-scale image–caption datasets, which are expensive and time-consuming to collect.

2.2 Reinforcement Learning

Increasingly, researchers adopt reinforcement learning to improve image captioning in LVLMs by optimizing task-specific reward signals. For instance, CLIP-based methods (Cho et al., 2022; Yu et al., 2023; Dzabraev et al., 2024) assess the correlation score between images and LVLM-generated captions based on Vision Language Models (VLMs) (Radford et al., 2021). Some approaches impose cycle-consistency by regenerating

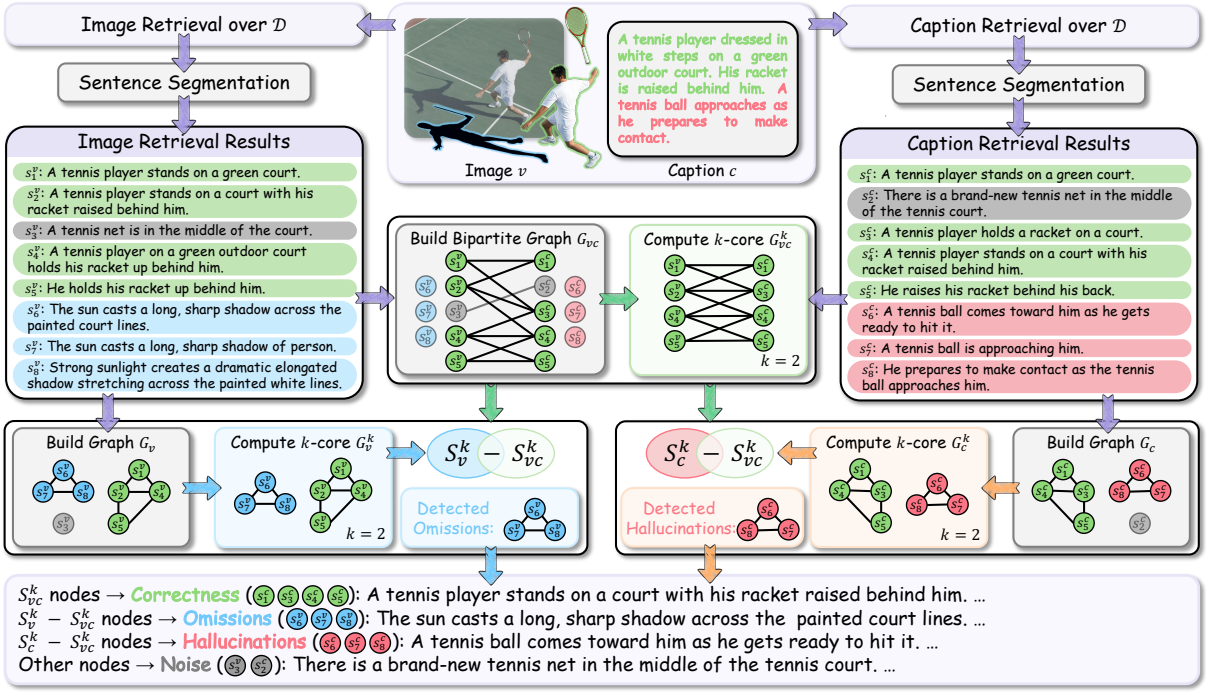

Figure 1: **Reasoning strategy of Caption Refinement Suggester (CRS) and Caption Quality Assessor (CQA).** We leverage k-core subgraph computation to analyze critical textual content in the retrieval results. Specifically, CRS extracts overlapping sentences S_v^k from the image retrieval results and uses them as signals to guide LVLMs to incorporate these key elements into their generated descriptions. By analyzing the discrepancy between S_v^k and caption-retrieved results S_c^k, CQA identifies hallucinations ($S_c^k - S_{vc}^k$) and omissions ($S_v^k - S_{vc}^k$) in this caption.

the image from the caption and using the reconstruction fidelity as the training signal (Feng et al., 2019; Bahng et al., 2025). Additionally, some methods use a self-retrieval objective, encouraging captions that can successfully retrieve originating images (Liu et al., 2018; Gaur et al., 2024; Dessì et al., 2023). Furthermore, reinforcement learning has been used to promote self-correction in captioning models (Zhang et al., 2025). Additionally, some work uses caption-conditioned downstream VQA accuracy as a reward signal (Xing et al., 2025). More recent work boosts the precision and detail richness of captions by minimizing information loss in modality conversion (Jia et al., 2026). However, RL-based methods still lag behind SFT.

3 Method

In this section, we introduce Retrieval-Guided Refinement for Image Captioning (Re^{3}Cap), a reasoning strategy to enhance the image captioning of LVLMs. Specifically, Caption Refinement Suggester (CRS) first identifies critical semantic elements within image captions. Subsequently, the Caption Quality Assessor (CQA) identifies omissions or misrepresentations in image captions. Leveraging the above guidance, our method finally encourages LVLMs to generate previously unexplored caption candidates.

3.1 Caption Refinement Suggester

Visually similar images often share overlapping semantic content. Based on this insight, the Caption Refinement Suggester (CRS) analyzes semantic elements that consistently appear across visually similar images. It then suggests LVLMs to incorporate these elements to refine their generated captions.

As shown in Figure 1, let v denote the image. The dataset $\mathcal{D} = \{p_i\}_{i=1}^N$ consists of N image-text pairs p_i, each containing an image x_i and its corresponding text t_i. With v as query, we perform image retrieval over the dataset $\mathcal{D}$ to obtain the top-K retrieval results denoted as $\mathcal{R}_v = \text{TopK}(\{\text{SIM}(v, x_i)\}_{i=1}^N) = \{p_i^v\}_{i=1}^K$. $\text{SIM}(v, x_i)$ means the correlation score between v and x_i calculated by the image retrieval model (Cherti et al., 2023). Through this process, we obtain a set of image-text pairs R_v. Each image in R_v shares similar visual representations to the query image v. Moreover, we construct a graph G_v to model descriptions corresponding to images in R_v. In G_v, each sentence is treated as a node. The textual

similarity score between every pair of nodes is computed by SBERT (Reimers and Gurevych, 2019). When the similarity score exceeds a predefined threshold τ, we establish an edge between these two nodes. Following algorithm (Seidman, 1983), we compute its k-core subgraph $G_v^k = (S_v^k, E_v^k)$ to analyze semantically consistent elements in G_v. E_v^k and S_v^k denote edges and nodes in the graph. By decomposing the k-core, CRS filters out long-tail descriptions in $\mathcal{R}_v$ and retains semantically consistent elements across image retrieval results.

In this way, our method leverages k-core analysis in visually similar images to identify semantic content (i.e., S_v^k) corresponding to the query image. During caption refinement, CRS suggests LVLMs to incorporate these descriptions, thereby improving the accuracy of the refined image caption.

3.2 Caption Quality Assessor

Semantically similar queries tend to yield consistent retrieval results. Motivated by this observation, the Caption Quality Assessor (CQA) evaluates discrepancies between the query image and its caption by comparing their respective retrieval results. By prompting LVLMs with the identified discrepancies between the image and its description, CQA guides LVLMs to generate more accurate captions.

Let c be the caption generated by the LVLM for the query image v, as illustrated in Figure 1. We then perform text retrieval over the dataset $\mathcal{D}$, using c as the query, to obtain the top-K retrieval results: $\mathcal{R}_c = \mathrm{TopK}(\{\mathrm{SIM}(c, t_i)\}_{i=1}^N) = \{p_i^c\}_{i=1}^K$. Similar to CRS, we construct a graph G_c over sentences in the caption-retrieved results $\mathcal{R}_c$ and compute its k-core subgraph $G_c^k = (S_c^k, E_c^k)$. Each node in G_c^k corresponds to a sentence that appears densely in caption retrieval results. We further construct a bipartite graph over $\mathcal{R}_v \cup \mathcal{R}_c$ and compute its k-core subgraph $G_{vc}^k = (S_{vc}^k, E_{vc}^k)$, which captures the semantically consistent content shared by the image and its corresponding caption. In this way, we can characterize the discrepancy between the image and its caption from two perspectives: hallucinated content in the caption is measured as $S_c^k - S_{vc}^k$, while $S_v^k - S_{vc}^k$ represents critical semantic content omitted by the LVLM in its generated caption.

Through information retrieval, CQA reformulates the complex cross-modal task of image caption quality assessment as an analysis of textual discrepancies within the retrieved results. As a result, the module can identify hallucinations and omissions in image captions without requiring additional annotations.

3.3 Overview of Re^{3}Cap

Based on the above reasoning strategy, we present the Retrieval-Guided Refinement for Image Captioning (Re^{3}Cap), a reinforcement learning framework that enhances image captioning without requiring additional annotations. As shown in Figure 2, given an input image v and a prompt q, we firstly sample a group of initial captions $\{c_i\}_{i=1}^M$ using LVLMs. With input image and initial captions as the queries, we perform image-conditioned retrieval and caption-conditioned retrieval. In this way, we leverage k-core analysis over retrieval results to generate guidance for caption improvement based on the Caption Refinement Suggester and the Caption Quality Assessor. The guidance $\{f_i\}_{i=1}^M$ identifies the correct elements, hallucinations, and omitted semantic content in LVLM-generated captions. Using the guidance, we further prompt the LVLMs to produce refined captions $\{c_i'\}_{i=1}^M$ that are more accurate and detailed. The reward function scores each refined caption with a reward R_i.

To enable the model to generate higher-quality captions directly from images during inference, Re^{3}Cap removes the initial caption and guidance used in the rollout stage from the policy input during policy optimization. This creates an off-policy optimization setting: the sampled captions are generated by a behavior policy conditioned on the image, the initial caption, and the guidance, whereas the optimized policy is conditioned only on the image. To correct for this distribution mismatch while still preserving a trust-region center for regularizing the policy update, we decouple the proximal policy from the behavior policy, following the decoupled PPO formulation (Hilton et al., 2022; Fu et al., 2026). Specifically, we optimize the following objective:

$$
\begin{aligned}
\mathcal{J}(\theta) = & \mathbb{E}_{v\sim\mathcal{V},\{c_i\}_{i=1}^M\sim\pi_{\theta_{\text{old}}}(\cdot|v),\{c_i'\}_{i=1}^M\sim\pi_{\theta_{\text{old}}}(\cdot|v,c_i,f_i)} \\
& \frac{1}{M}\sum_{i=1}^{M}\frac{1}{|c_i'|}\sum_{t=1}^{|c_i'|}(\min(\frac{\pi_\theta}{\pi_{\text{behav}}}\hat{A}_{i,t}, \\
& \frac{\pi_{\text{prox}}}{\pi_{\text{behav}}}\,\text{clip}(\frac{\pi_\theta}{\pi_{\text{prox}}}, 1-\epsilon, 1+\epsilon)\hat{A}_{i,t}) \\
& - \beta D_{\text{KL}}(\pi_\theta \| \pi_{\text{ref}})),
\end{aligned}
$$

where

$$
\begin{aligned}
\pi_{\text{behav}} &= \pi_{\theta_{\text{old}}}\left(c_{i,t}' \mid v, q, c_i, f_i, c_{i,<t}'\right), \\
\pi_{\text{prox}} &= \pi_{\theta_{\text{old}}}\left(c_{i,t}' \mid v, q, c_{i,<t}'\right), \\
\pi_\theta &= \pi_\theta\left(c_{i,t}' \mid v, q, c_{i,<t}'\right).
\end{aligned}
$$

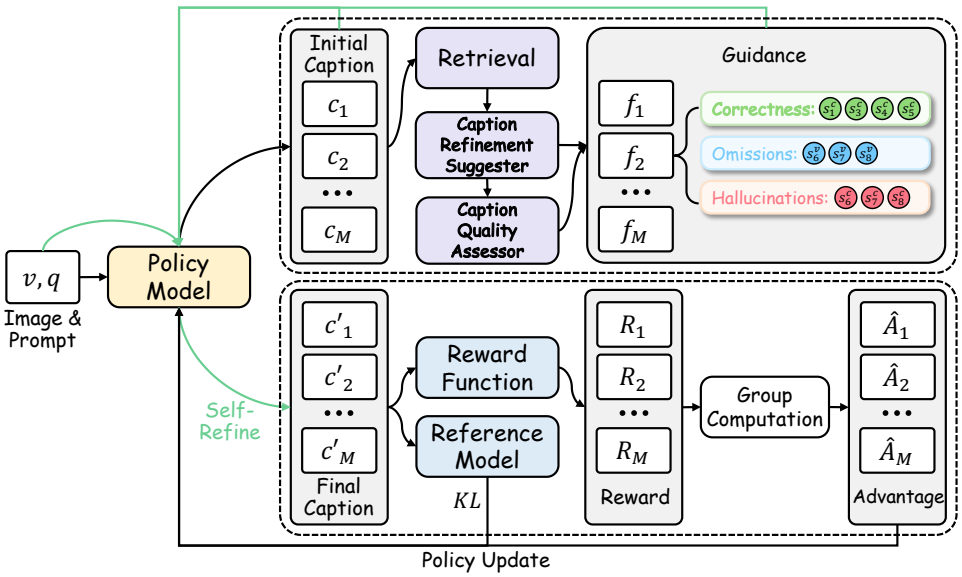

Figure 2: **Overview of Retrieval-Guided Refinement for Image Captioning (Re^{3}Cap).** Re^{3}Cap begins by sampling initial captions $\{c_i\}_{i=1}^M$ for each image v. Next, it performs image-conditioned retrieval and caption-conditioned retrieval, and leverages k-core analysis to generate guidance $\{f_i\}_{i=1}^M$ based on Caption Refinement Suggester and Caption Quality Assessor. Finally, the method incorporates the guidance into prompts to obtain refined captions $\{c'_i\}_{i=1}^M$ and optimizes the policy model with reinforcement learning.

As a result, the optimized policy is conditioned solely on the image, eliminating the need for retrieval or k-core analysis at inference time.

4 Experiment

In this section, we first introduce our experimental settings. We then present a reasoning capability analysis. Subsequently, we demonstrate the effectiveness of our method by comparing it with GRPO and state-of-the-art image captioning methods. Finally, we present an ablation study to investigate the contribution of each component. Additional experiments, including robustness across diverse encoders, sensitivity to hyperparameter choices such as the retrieval number K and the similarity threshold τ, and computational overhead, are provided in the Section B.

4.1 Experimental Settings

Training settings. Following (Zhang et al., 2025) and CIM (Jia et al., 2026), we use images from the RefinedCaps dataset (Zhang et al., 2025) as the training set, consisting of 6.5K images sampled from the COCO training split (Lin et al., 2014). For both image-conditioned and caption-conditioned retrieval, we retrieve the top-K candidates ($K = 3$) from a dataset constructed by augmenting RefinedCaps (Zhang et al., 2025) with DenseFusion-1M (Li et al., 2024b). To avoid data leakage, we ensure that the retrieval corpus is disjoint from all evaluation benchmarks. We use SBERT (Reimers and Gurevych, 2019) with MPNet-base backbone (Song et al., 2020) as the text encoder and OpenCLIP ViT-H/14 (Cherti et al., 2023) as the image encoder, respectively. For CRS and CQA, we set the k in the k-core to $k = \lceil K/2 \rceil = 2$, and use a threshold $\tau = 0.7$. We adopt the VERL framework (Sheng et al., 2025) for training. For hyperparameters, we utilize the Adam optimizer and train for two epochs with a constant learning rate of 1×10^{-6}. For rollout, the prompt batch size is 256, and we sample $M = 5$ responses for each prompt. For training, the mini-batch size is set to 64. We set the clipping ratio to $\epsilon = 0.2$ and the KL penalty coefficient to $\beta = 0.001$.

Models. To validate the generalizability of Re^{3}Cap, we evaluate its performance across representative Large Vision-Language Models (LVLMs): LLaVA-1.5-7B (Liu et al., 2023), Qwen2-VL-7B (Wang et al., 2024b), and Qwen2.5-VL-7B (Bai et al., 2025b). Additional evaluations on InternVL3-8B (Zhu et al., 2025) and Qwen3-VL-8B (Bai et al., 2025a) are provided in Section B.1.

Benchmarks. We use COCO-LN500 (Pont-Tuset et al., 2020) and DOCCI500 (Onoe et al., 2024) as the evaluation benchmarks to validate the effectiveness of our proposed Re^{3}Cap. COCO-LN500 (Pont-Tuset et al., 2020) consists

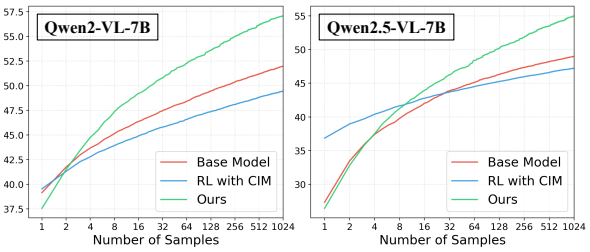

Figure 3: **Capability boundary analysis with max@k.**

of 500 image–caption pairs from the Localized-narratives test set in COCO2017 (Lin et al., 2014). DOCCI500 (Onoe et al., 2024) is a random sample of 500 image-caption pairs from DOCCI test split, where images largely lack human-centric content.

Metrics. Following SC-Captioner (Zhang et al., 2025), we evaluate caption quality using the F1 score from three aspects: objects, attributes, and relations. For relations, we measure relational correctness via VQA-based accuracy based on Qwen3 (Yang et al., 2025a).

Baselines. We adopt Group Relative Policy Optimization (GRPO) (Shao et al., 2024) as the baseline with multiple reward functions. We consider CLIP (Cho et al., 2022), which uses the CLIP (Radford et al., 2021) image-text similarity score as the reward; SC (Zhang et al., 2025), which rewards keyword-level self-correction; and CIM (Jia et al., 2026), which rewards the similarity between images retrieved by the caption and original image.

4.2 Reasoning Capability Analysis

Inspired by Yue et al. (2025), we further evaluate whether our retrieval-guided reasoning strategy enables the LVLM to explore caption candidates beyond those already covered by the base model. Specifically, for each image, we sample multiple candidate captions from each model and evaluate them on COCO-LN500 (Pont-Tuset et al., 2020) using the BLEU-4 (Papineni et al., 2002) score. Since BLEU-4 is a continuous metric, we adopt max@k, a continuous generalization of pass@k (Bagirov et al., 2025), to measure the best achievable caption quality under a large sampling budget. We compute max@k using the unbiased low-variance estimator proposed by Walder and Karkhanis (2026).

As shown in Figure 3, the reinforcement learning method using CIM (Jia et al., 2026) as the reward function achieves strong performance at $k = 1$, indicating that conventional RL effectively improves sampling efficiency. However, as k increases, it grows more slowly and eventually falls below the base model, suggesting that it tends to narrow the output distribution and does not preserve sufficient exploration diversity. In contrast, our retrieval-guided reasoning strategy, without any training, consistently benefits from larger sampling budgets and surpasses the base model as k increases. This trend indicates that our method encourages the model to generate more diverse and previously unexplored caption candidates. Therefore, the results demonstrate that our method can expand the capability boundary of the base model. The experiments in Section 4.3 further validate this conclusion by showing that, when incorporated into reinforcement learning, our strategy brings consistent improvements across different LVLMs, benchmarks, and reward functions.

4.3 Reinforcement Learning on Base Model

We verify the generalization of our method via various LVLMs and evaluate performance on benchmarks (Pont-Tuset et al., 2020; Onoe et al., 2024). In tables, Base means the LVLMs without any task-specific training, and SFT denotes the LVLMs supervised fine-tuned on the Refined-Caps dataset (Zhang et al., 2025). GRPO and Ours denote models trained from the base model with GRPO and Re^{3}Cap, respectively.

As shown in Table 1, the results demonstrate that our method consistently outperforms GRPO across multiple LVLMs and with various reward functions, especially on the more challenging relation reasoning task. For instance, on the QA score of the Relations evaluation, our method achieves improvements by 8.64% on COCO-LN500 (Pont-Tuset et al., 2020) and 7.57% on DOCCI500 (Onoe et al., 2024), averaged across multiple base LVLMs and reward functions. Moreover, the gains are particularly pronounced when using weaker base LVLMs and reward functions. With CLIP (Cho et al., 2022) as the reward function, our method achieves gains of 4.39% in Objects F1, 2.44% in Attributes F1, and 6.74% in Relations QA, averaged across multiple base LVLMs and both benchmarks. For LLaVA1.5-7B (Liu et al., 2023) as the base LVLM, our method yields gains of 5.91% in Objects F1, 2.06% in Attributes F1, and 8.95% in Relations QA, averaged across multiple reward functions and both benchmarks. Notably, using LLaVA1.5-7B (Liu et al., 2023) as the base LVLM, our method even outperforms SFT across multiple reward functions and both benchmarks, whereas GRPO underperforms SFT. This demonstrates that GRPO is constrained

Base Model	Reward	Method	COCO-LN500			DOCCI500		
			Objects F1	Attributes F1	Relations QA	Objects F1	Attributes F1	Relations QA
LLaVA1.5 -7B	–	○ Base	67.56	42.34	14.38	59.41	48.01	9.19
		○ SFT	73.45	54.25	28.59	68.44	53.93	19.87
	CLIP	○ GRPO	66.66	53.15	22.58	61.77	53.62	18.42
		● Ours	**72.03** (↑5.4)	**54.73** (↑1.6)	**30.38** (↑7.8)	**68.51** (↑6.7)	**57.68** (↑4.1)	**23.86** (↑5.4)
	SC	○ GRPO	69.49	50.44	21.77	62.52	53.80	17.86
		● Ours	**74.08** (↑4.6)	**54.57** (↑4.1)	**35.17** (↑13.4)	**70.25** (↑7.7)	**56.21** (↑2.4)	**26.92** (↑9.1)
	CIM	○ GRPO	69.80	54.38	24.98	63.38	**56.28**	19.87
		● Ours	**74.74** (↑4.9)	**54.73** (↑0.4)	**34.93** (↑10.0)	**69.49** (↑6.1)	56.13 (↓0.2)	**27.93** (↑8.1)
Qwen2 -VL-7B	–	○ Base	69.47	48.68	20.47	66.47	52.65	17.57
		○ SFT	75.37	56.54	36.39	69.50	55.50	27.65
	CLIP	○ GRPO	69.04	53.37	26.48	67.30	54.63	26.28
		● Ours	**75.33** (↑6.3)	**58.35** (↑5.0)	**35.26** (↑8.8)	**71.88** (↑4.6)	**58.37** (↑3.7)	**32.73** (↑6.5)
	SC	○ GRPO	76.80	57.49	30.46	72.49	57.75	23.50
		● Ours	**77.77** (↑1.0)	**58.79** (↑1.3)	**44.60** (↑14.1)	**73.29** (↑0.8)	**58.88** (↑1.1)	**36.40** (↑12.9)
	CIM	○ GRPO	75.80	58.22	38.71	71.43	**59.18**	32.12
		● Ours	**78.18** (↑2.4)	**59.28** (↑1.1)	**44.19** (↑5.5)	**73.51** (↑2.1)	58.99 (↓0.2)	**40.15** (↑8.0)
Qwen2.5 -VL-7B	–	○ Base	65.37	46.25	23.76	65.06	52.27	24.35
		○ SFT	75.72	57.09	39.64	71.94	58.28	34.38
	CLIP	○ GRPO	68.32	**53.90**	23.80	68.21	55.32	24.79
		● Ours	**70.52** (↑2.2)	53.80 (↓0.1)	**30.26** (↑6.5)	**69.39** (↑1.2)	**55.70** (↑0.4)	**30.27** (↑5.5)
	SC	○ GRPO	**77.52**	56.71	31.19	**72.81**	57.93	28.01
		● Ours	76.85 (↓0.7)	**58.96** (↑2.3)	**41.06** (↑9.9)	**72.81** (↑0.0)	**58.51** (↑0.6)	**36.80** (↑8.8)
	CIM	○ GRPO	77.59	58.51	44.15	71.88	59.08	34.70
		● Ours	**77.80** (↑0.2)	**59.26** (↑0.8)	**46.02** (↑1.9)	**72.80** (↑0.9)	**59.77** (↑0.7)	**38.65** (↑4.0)

Table 1: **Performance comparison of Reinforcement Learning with GRPO and Re^{3}Cap across multiple reward functions on COCO-LN500 (Pont-Tuset et al., 2020) and DOCCI500 (Onoe et al., 2024).** Constrained by the reasoning capacity of base models, GRPO struggles to surpass the performance of task-specific SFT models, especially on weaker LVLMs such as LLaVA-1.5-7B (Liu et al., 2023). In contrast, our method introduces a novel reasoning strategy to generate unexplored caption candidates, consistently improving image captioning performance across weaker and stronger base models, including Qwen2-VL-7B (Wang et al., 2024b) and Qwen2.5-VL-7B (Bai et al., 2025b). Moreover, using a more accurate reward further improves the performance of ours.

by the reasoning capacity of base models, making it difficult to surpass the performance of models supervised fine-tuned on task-specific datasets. In contrast, our method can generate previously unexplored caption candidates from base LVLMs by injecting a novel reasoning strategy, leading to substantial gains. For Qwen2-VL-7B (Wang et al., 2024b) and Qwen2.5-VL-7B (Bai et al., 2025b) as the base LVLMs, GRPO can surpass SFT when using CIM (Jia et al., 2026) as the reward function. Even on these strong base LVLMs, our method further improves performance, indicating its effectiveness. The above results demonstrate that Re^{3}Cap significantly enhances the image captioning capability of LVLMs with distinct architectures.

4.4 SOTA Comparison

To further verify the superiority of our method, we conduct experiments to compare it with state-of-the-art methods. As shown in Table 2, VCD (Leng et al., 2024) and INTER (Dong et al., 2025) are training-free methods designed to mitigate hallucinations in LVLMs. SC-Captioner (Zhang et al., 2025) and SFT+CIM (Jia et al., 2026) both perform reinforcement learning with their respective reward functions after first applying Supervised Fine-Tuning (SFT) to the base LVLMs. In contrast, Ours refers to the base model solely optimized by Re^{3}Cap via reinforcement learning, without supervised fine-tuning on task-specific datasets.

The results demonstrate that our approach achieves superior performance in image captioning, even compared with SFT-based methods. Specifically, on the more challenging relation reasoning task, our method improves Relations QA by 4.08% on COCO-LN500 (Pont-Tuset et al., 2020) and 4.86% on DOCCI500 (Onoe et al., 2024), averaged across multiple reward functions. Additionally, our

Base Model	Reward	Method	COCO-LN500			DOCCI500		
			Objects F1	Attributes F1	Relations QA	Objects F1	Attributes F1	Relations QA
Qwen2 -VL-7B	–	○ VCD	69.85	48.71	26.77	67.57	53.80	25.03
		○ INTER	69.81	48.58	26.93	67.42	53.70	24.67
	SC	○ SC-Captioner	76.37	57.56	38.51	71.63	57.67	30.51
		• Ours	**77.77** (↑1.4)	**58.79** (↑1.2)	**44.60** (↑6.1)	**73.29** (↑1.7)	**58.88** (↑1.2)	**36.40** (↑5.9)
	CIM	○ SFT+CIM	76.65	58.09	42.12	73.87	58.68	36.32
		• Ours	**78.18** (↑1.5)	**59.28** (↑1.2)	**44.19** (↑2.1)	**73.92** (↑0.1)	**58.99** (↑0.3)	**40.15** (↑3.8)

Table 2: **Performance comparison with state-of-the-art methods on Qwen2-VL-7B (Wang et al., 2024b) over COCO-LN500 (Pont-Tuset et al., 2020) and DOCCI500 (Onoe et al., 2024).** The results demonstrate the superiority of our proposed reinforcement learning framework when compared with state-of-the-art methods.

method achieves gains of 1.47% in Objects F1 and 1.21% in Attributes F1 on COCO-LN500 (Pont-Tuset et al., 2020), averaged across multiple reward functions. Notably, our method achieves these improvements with only a single-stage RL training, whereas SC-Captioner and SFT+CIM rely on a two-stage pipeline (SFT followed by RL). Moreover, our method outperforms the training-free methods across all metrics by a large margin. Such results indicate that by injecting the reasoning strategy during reinforcement learning, our method achieves superior performance in image captioning.

4.5 Ablation Study

In this section, we present an ablation study to quantitatively evaluate the effectiveness of each core component (CRS and CQA) within our framework. As shown in Table 3, the first row denotes the model trained solely with GRPO (Shao et al., 2024). The middle two rows refer to models optimized by reinforcement learning that use CRS and CQA as their reasoning strategy, respectively. The last row denotes the model trained using Re^{3}Cap, combined with CRS and CQA.

The results in Table 3 show that each core component achieves consistent performance improvements. Specifically, CRS achieves improvements of 1.03% in Objects F1, 0.76% in Attributes F1, and 2.19% in Relations QA on COCO-LN500. Moreover, CQA achieves improvements of 1.34% in Objects F1, 0.82% in Attributes F1, and 3.98% in Relations QA. Such results indicate CRS improves performance by guiding the LVLM to retain accurate descriptions, and CQA guides the LVLM to mitigate hallucinations and reduce omissions, thereby further improving performance. Most importantly, our method achieves the best performance by combining the complementary CRS and CQA.

Components		Objects	Attributes	Relations
CRS	CQA	F1	F1	QA
✗	✗	75.80	58.22	38.71
✓	✗	76.83	58.98	40.90
✗	✓	77.14	59.04	42.69
✓	✓	**78.18**	**59.28**	**44.19**

Table 3: **Ablation study of core components on COCO-LN500 (Pont-Tuset et al., 2020) using Qwen2-VL-7B (Wang et al., 2024b) with CIM (Jia et al., 2026) as the reward function.**

5 Conclusion

In this paper, we present Re^{3}Cap, a reinforcement learning framework that consistently outperforms previous Supervised Fine-Tuning (SFT) approaches in image captioning. Our key observation is that discrepancies between retrieval results reveal potential hallucinations and omitted visual details in image captions, providing an informative signal for assessing caption quality. Building on this insight, we further propose a retrieval-based reasoning strategy that guides LVLMs to generate previously unexplored caption candidates. By performing reinforcement learning on these newly explored candidates, the model effectively expands the caption space and refines its generation behavior. Extensive experiments demonstrate that our proposed Re^{3}Cap enables LVLMs to achieve consistently superior performance in image captioning, even compared with strong Supervised Fine-Tuning (SFT) baselines. Overall, this work enhances the reasoning capabilities of LVLMs in reinforcement learning and offers a new perspective on improving model performance in image captioning. We hope the proposed framework provides more insights for future research in both multimodal reasoning and caption generation.

Limitations

Specifically, the effectiveness of our method depends on the quality and scale of the retrieval set. When the dataset is too small, many image captions fail to retrieve relevant results. In such cases, our method may degenerate into a simple reinforcement learning approach. We believe that increasing the scale and diversity of the retrieval set would improve the robustness of our approach.

Ethical Considerations

This work aims to improve image captioning by introducing a retrieval-guided refinement strategy during reinforcement learning. All experiments are conducted on publicly available image-caption datasets and benchmarks. As in prior work, these datasets and evaluation protocols may contain social biases, annotation artifacts, sampling biases, or other imperfections that can affect model behavior and evaluation outcomes. Beyond the risks already associated with multimodal model training, retrieval, and evaluation on existing public datasets, we do not identify additional ethical risks introduced specifically by our method.

References

Farid Bagirov, Mikhail Arkhipov, Ksenia Sycheva, Evgeniy Glukhov, and Egor Bogomolov. 2025. The best of n worlds: Aligning reinforcement learning with best-of-n sampling via max@ k optimisation. *arXiv preprint arXiv:2510.23393*.

Hyojin Bahng, Caroline Chan, Fredo Durand, and Phillip Isola. 2025. Cycle consistency as reward: Learning image-text alignment without human preferences. *arXiv preprint arXiv:2506.02095*.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*.

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

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025b. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, and 1 others. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8.

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

Yuxin Chen, Zongyang Ma, Ziqi Zhang, Zhongang Qi, Chunfeng Yuan, Bing Li, Junfu Pu, Ying Shan, Xiaojuan Qi, and Weiming Hu. 2024b. How to make cross encoder a good teacher for efficient image-text retrieval? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26994–27003.

Xianfu Cheng, Wei Zhang, Shiwei Zhang, Jian Yang, Xiangyuan Guan, Xianjie Wu, Xiang Li, Ge Zhang, Jiaheng Liu, Yuying Mai, and 1 others. 2025. Simplevqa: Multimodal factuality evaluation for multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4637–4646.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2818–2829.

Jaemin Cho, Seunghyun Yoon, Ajinkya Kale, Franck Dernoncourt, Trung Bui, and Mohit Bansal. 2022. Fine-grained image captioning with clip reward. *arXiv preprint arXiv:2205.13115*.

Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. 2020. Meshed-memory transformer for image captioning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10578–10587.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Preprint*, arXiv:2305.06500.

Roberto Dessì, Michele Bevilacqua, Eleonora Gualdoni, Nathanaël Carraz Rakotonirina, Francesca Franzon, and Marco Baroni. 2023. Cross-domain image captioning with discriminative finetuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6935–6944.

Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Bin Wang, Linke Ouyang, Xilin Wei, Songyang Zhang, Haodong Duan, Maosong Cao, and 1 others. 2024. Internlm-xcomposer2: Mastering free-form text-image composition and comprehension in vision-language large model. *arXiv preprint arXiv:2401.16420*.

Xin Dong, Shichao Dong, Jin Wang, Jing Huang, Li Zhou, Zenghui Sun, Lihua Jing, Jinsong Lan, Xiaoyong Zhu, and Bo Zheng. 2025. Inter: Mitigating hallucination in large vision-language models by interaction guidance sampling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2534–2544.

Siyuan Duan, Yuan Sun, Dezhong Peng, Zheng Liu, Xiaomin Song, and Peng Hu. 2025. Fuzzy multimodal learning for trusted cross-modal retrieval. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 20747–20756.

Maksim Dzabraev, Alexander Kunitsyn, and Andrei Ivaniuta. 2024. Vlrm: Vision-language models act as reward models for image captioning. *arXiv preprint arXiv:2404.01911*.

Junjie Fei, Teng Wang, Jinrui Zhang, Zhenyu He, Chengjie Wang, and Feng Zheng. 2023. Transferable decoding with visual entities for zero-shot image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3136–3146.

Yang Feng, Lin Ma, Wei Liu, and Jiebo Luo. 2019. Unsupervised image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4125–4134.

Wei Fu, Jiaxuan Gao, Xujie Shen, Chen Zhu, Zhiyu Mei, Chuyi He, Shusheng Xu, Guo Wei, Jun Mei, Jiashu Wang, and 1 others. 2026. Areal: A large-scale asynchronous reinforcement learning system for language reasoning. *Advances in Neural Information Processing Systems*, 38:36256–36282.

Manu Gaur, Darshan Singh, and Makarand Tapaswi. 2024. No detail left behind: Revisiting self-retrieval for fine-grained image captioning. *arXiv preprint arXiv:2409.03025*.

Jacob Hilton, Karl Cobbe, and John Schulman. 2022. Batch size-invariance for policy optimization. *Advances in Neural Information Processing Systems*, 35:17086–17098.

Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. 2024. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22170–22183.

Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. 2019. Attention on attention for image captioning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4634–4643.

Haonan Jia, Shichao Dong, Xin Dong, Zenghui Sun, Jin Wang, Jinsong Lan, Xiaoyong Zhu, Bo Zheng, and Kaifu Zhang. 2026. Cross-modal identity mapping: Minimizing information loss in modality conversion via reinforcement learning. *arXiv preprint arXiv:2603.01696*.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

Taewhan Kim, Soeun Lee, Si-Woo Kim, and Dong-Jin Kim. 2025. Vipcap: Retrieval text-based visual prompts for lightweight image captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 4320–4328.

Simon Kornblith, Lala Li, Zirui Wang, and Thao Nguyen. 2023. Guiding image captioning models toward more specific captions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15259–15269.

Jeong Ryong Lee, Yejee Shin, Geonhui Son, and Dosik Hwang. 2025. Diffusion bridge: Leveraging diffusion model to reduce the modality gap between text and vision for zero-shot image captioning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4050–4059.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882.

Jiaxuan Li, Duc Minh Vo, Akihiro Sugimoto, and Hideki Nakayama. 2024a. Evcap: Retrieval-augmented image captioning with external visual-name memory for open-world comprehension. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13733–13742.

Xiaotong Li, Fan Zhang, Haiwen Diao, Yueze Wang, Xinlong Wang, and Lingyu Duan. 2024b. Densefusion-1m: Merging vision experts for comprehensive multimodal perception. *Advances in Neural Information Processing Systems*, 37:18535–18556.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Chenxi Liu, Junhua Mao, Fei Sha, and Alan Yuille. 2017a. Attention correctness in neural image captioning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning.

Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. 2017b. Improved image captioning via policy gradient optimization of spider. In *Proceedings of the IEEE international conference on computer vision*, pages 873–881.

Xihui Liu, Hongsheng Li, Jing Shao, Dapeng Chen, and Xiaogang Wang. 2018. Show, tell and discriminate: Image captioning by self-retrieval with partially labeled data. In *Proceedings of the European conference on computer vision (ECCV)*, pages 338–354.

Jianjie Luo, Jingwen Chen, Yehao Li, Yingwei Pan, Jianlin Feng, Hongyang Chao, and Ting Yao. 2024. Unleashing text-to-image diffusion prior for zero-shot image captioning. In *European Conference on Computer Vision*, pages 237–254. Springer.

Ziyang Luo, Zhipeng Hu, Yadong Xi, Rongsheng Zhang, and Jing Ma. 2023. I-tuning: Tuning frozen language models with image for lightweight image captioning. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

Ziqi Miao, Haonan Jia, Lijun Li, Chen Qian, Yuan Xiong, Wenting Yan, and Jing Shao. 2026. Seeing with you: Perception-reasoning coevolution for multimodal reasoning. *arXiv preprint arXiv:2603.28618*.

Ron Mokady, Amir Hertz, and Amit H Bermano. 2021. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*.

Yasumasa Onoe, Sunayana Rane, Zachary Berger, Yonatan Bitton, Jaemin Cho, Roopal Garg, Alexander Ku, Zarana Parekh, Jordi Pont-Tuset, Garrett Tanzer, and 1 others. 2024. Docci: Descriptions of connected and contrasting images. In *European Conference on Computer Vision*, pages 291–309. Springer.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. 2020. Connecting vision and language with localized narratives. In *European conference on computer vision*, pages 647–664. Springer.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

Rita Ramos, Bruno Martins, Desmond Elliott, and Yova Kementchedjhieva. 2023. Smallcap: lightweight image captioning prompted with retrieval augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2840–2849.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Kuniaki Saito, Donghyun Kim, Kwanyong Park, Atsushi Hashimoto, and Yoshitaka Ushiku. 2025. Captionsmiths: Flexibly controlling language pattern in image captioning. *arXiv preprint arXiv:2507.01409*.

Stephen B Seidman. 1983. Network structure and minimum degree. *Social networks*, 5(3):269–287.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. 2025. Hybridflow: A flexible and efficient rlhf framework. In *Proceedings of the Twentieth European Conference on Computer Systems*, pages 1279–1297.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, and 1 others. 2025. Dinov3. *arXiv preprint arXiv:2508.10104*.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33:16857–16867.

Derek Tam, Colin Raffel, and Mohit Bansal. 2023. Simple weakly-supervised image captioning via clip's multimodal embeddings. In *The AAAI-23 Workshop on Creative AI Across Modalities*.

Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. 2022. Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17918–17928.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164.

Christian Walder and Deep Tejas Karkhanis. 2026. Pass@ k policy optimization: Solving harder reinforcement learning problems. *Advances in Neural Information Processing Systems*, 38:152416–152445.

Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. 2022. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100*.

Jin Wang, Shichao Dong, Yapeng Zhu, Kelu Yao, Weidong Zhao, Chao Li, and Ping Luo. 2024a. Diagnosing the compositional knowledge of vision language models from a game-theoretic view. *arXiv preprint arXiv:2405.17201*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in neural information processing systems*, 33:5776–5788.

Long Xing, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Jianze Liang, Qidong Huang, Jiaqi Wang, Feng Wu, and Dahua Lin. 2025. Caprl: Stimulating dense image caption capabilities via reinforcement learning. *arXiv preprint arXiv:2509.22647*.

Dongsheng Xu, Wenye Zhao, Yi Cai, and Qingbao Huang. 2023. Zero-textcap: Zero-shot framework for text-based image captioning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4949–4957.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Zhantao Yang, Ruili Feng, Keyu Yan, Huangji Wang, Zhicai Wang, Shangwen Zhu, Han Zhang, Jie Xiao, Pingyu Wu, Kai Zhu, and 1 others. 2025b. Bacon: Improving clarity of image captions via bag-of-concept graphs. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 14380–14389.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Jiarui Yu, Haoran Li, Yanbin Hao, Bin Zhu, Tong Xu, and Xiangnan He. 2023. Cgt-gan: Clip-guided text gan for image captioning. In *Proceedings of the 31st ACM international conference on multimedia*, pages 2252–2263.

Yang Yue, Zhiqi Chen, Rui Lu, Andrew Zhao, Zhaokai Wang, Shiji Song, and Gao Huang. 2025. Does reinforcement learning really incentivize reasoning capacity in llms beyond the base model? *arXiv preprint arXiv:2504.13837*.

Lin Zhang, Xianfang Zeng, Kangcong Li, Gang Yu, and Tao Chen. 2025. Sc-captioner: Improving image captioning with self-correction by reinforcement learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23145–23155.

Wendi Zheng, Jiayan Teng, Zhuoyi Yang, Weihan Wang, Jidong Chen, Xiaotao Gu, Yuxiao Dong, Ming Ding, and Jie Tang. 2024. Cogview3: Finer and faster text-to-image generation via relay diffusion. In *European Conference on Computer Vision*, pages 1–22. Springer.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

A Prompt Templates

We use a concise prompt to generate the initial caption. For Re^{3}Cap, the model analyzes the initial caption with the reasoning strategy and produces guidance, which is then injected into the response to prompt the model to generate a refined caption. For Relation QA, we prompt Qwen3 (Yang et al., 2025a) models to answer the given questions based on the candidate captions. The detailed prompts are shown in Figure 4.

B Additional Experiments

B.1 Reinforcement Learning on More LVLMs

We further adopt InternVL3-8B (Zhu et al., 2025) and Qwen3-VL-8B (Bai et al., 2025a) as the base LVLM to validate the effectiveness of our method. As shown in Table 4, the results demonstrate that our method consistently outperforms GRPO across multiple reward functions, especially on the more challenging relation reasoning task. Specifically, for InternVL3-8B, Re^{3}Cap improves the Relations QA score over GRPO by an average of 6.59% on COCO-LN500 (Pont-Tuset et al., 2020) and 5.12% on DOCCI500 (Onoe et al., 2024) across multiple reward functions. For Qwen3-VL-8B, Re^{3}Cap also brings consistent improvements, achieving average gains of 3.76% on COCO-LN500 and 2.42% on DOCCI500 in Relations QA over GRPO. Moreover, GRPO fails to surpass SFT under the CIM (Jia et al., 2026) and SC (Zhang et al., 2025) reward functions, whereas our method consistently outperforms SFT under these reward signals. The above results demonstrate that our proposed Re^{3}Cap significantly enhances the image captioning capability of LVLMs.

B.2 Robustness of the Choice of Hyperparameters

We conduct experiments to evaluate the effects of two hyperparameters on model performance: the retrieval number K and the similarity threshold τ used for edge construction. As shown in Figure 5 and Figure 6, we vary K from 3 to 11 and τ from 0.5 to 0.9. The results show only minor performance variations across different settings. Specifically, on COCO-LN500 (Pont-Tuset et al., 2020) with Qwen2-VL-7B (Wang et al., 2024b), the variations in Objects F1, Attributes F1, and Relation QA are within 0.53%, 0.74%, and 1.44%, respectively, across different K values, and within 1.32%, 1.48%, and 1.38%, respectively, across different τ values. Overall, such results demonstrate that our method is robust to the choice of hyperparameters.

B.3 Robustness across Diverse Encoders

We conduct experiments to evaluate the impact of different encoders used in our method for retrieval and graph edge construction. As shown in Table 5, we use either DINOv3 ViT-L/16 (Siméoni et al., 2025) or OpenCLIP ViT-H/14 (Cherti et al., 2023) as the image encoder, and SBERT (Reimers and Gurevych, 2019) with a MiniLM-base (Wang et al., 2020) or MPNet-base (Song et al., 2020) backbone as the text encoder. The results show only minor performance variations across different encoders. Specifically, the variations in Objects F1, Attributes F1, and Relation QA are within 0.46%, 0.79%, and 0.94%, respectively, indicating strong robustness to the choice of both image and text encoders. Overall, such results demonstrate that our method is robust to the potential information loss and representation discrepancies introduced by different encoders.

B.4 Analysis of Computational Overhead

Taking SC (Zhang et al., 2025) as the reward function as an example, GRPO requires 192 GPU hours on the NVIDIA A100 to converge, whereas our method converges in approximately 216 GPU hours under the same experimental settings. This corresponds to only an additional 24 GPU hours, or about 12.5% more training time, while achieving significant performance gains. Moreover, the memory overhead of our method is comparable to that of GRPO.

Prompt for Initial Caption

User: Caption this image as accurately as possible, without speculation. Describe what you see.
Assistant:

Prompt for Re^{3}Cap

User: Caption this image as accurately as possible, without speculation. Describe what you see.
Assistant: <Initial Caption>
Wait. Let me check this caption again against the image.
I should KEEP: <Correctness>
I should ADD: <Omissions>
I should REMOVE: <Hallucinations>
Final caption:

Prompt for Relation QA

User: I will give you a passage of caption. Please answer the following 5 questions with "Yes", "No", or "n/a" based on the given caption. Output like this: 1: Yes, 2: No, 3: Yes, 4: n/a, 5: Yes. Don't output extra text.
Caption: <Caption>
Questions:1.<Question1> 2.<Question2> 3.<Question3> 4.<Question4> 5.<Question5>
Assistant:

Figure 4: Prompt example for initial caption, Re^{3}Cap, and relation evaluation.

Base Model	Reward	Method	COCO-LN500			DOCCI500		
			Objects F1	Attributes F1	Relations QA	Objects F1	Attributes F1	Relations QA
InternVL3-8B	–	○ Base	71.00	50.66	26.44	66.08	53.72	25.11
		○ SFT	76.42	57.79	42.32	72.72	58.59	35.31
	CLIP	○ GRPO	72.14	**54.84**	30.83	68.70	**57.58**	28.13
		● Ours	**74.79** (↑2.7)	54.42 (↓0.4)	**35.46** (↑4.6)	**71.15** (↑2.5)	54.85 (↓2.7)	**28.81** (↑0.7)
	SC	○ GRPO	75.77	57.14	35.58	70.72	57.78	28.38
		● Ours	**78.51** (↑2.7)	**59.50** (↑2.4)	**45.21** (↑9.6)	**74.11** (↑3.4)	**58.19** (↑0.4)	**36.84** (↑8.5)
	CIM	○ GRPO	76.14	58.70	38.67	70.47	59.26	30.39
		● Ours	**78.68** (↑2.5)	**58.99** (↑0.3)	**44.19** (↑5.5)	**73.34** (↑2.9)	**59.78** (↑0.5)	**36.60** (↑6.2)
Qwen3-VL-8B	–	○ Base	72.64	53.73	37.57	72.68	56.08	40.67
		○ SFT	76.59	57.14	42.24	72.51	57.37	38.65
	CLIP	○ GRPO	73.91	54.72	39.84	72.93	56.31	39.82
		● Ours	**74.86** (↑1.0)	**55.83** (↑1.1)	**41.37** (↑1.5)	**73.24** (↑0.3)	**56.73** (↑0.4)	**41.36** (↑1.5)
	SC	○ GRPO	75.11	54.98	42.85	72.96	56.64	40.67
		● Ours	**76.45** (↑1.3)	**57.31** (↑2.3)	**46.26** (↑3.4)	**73.76** (↑0.8)	**57.02** (↑0.4)	**41.92** (↑1.3)
	CIM	○ GRPO	75.21	56.03	39.52	72.27	57.39	38.17
		● Ours	**76.92** (↑1.7)	**57.98** (↑2.0)	**45.86** (↑6.3)	**73.60** (↑1.3)	**57.94** (↑0.6)	**42.64** (↑4.5)

Table 4: **Performance comparison of reinforcement learning with GRPO and Re^{3}Cap on more LVLMs.** We evaluate InternVL3-8B (Zhu et al., 2025) and Qwen3-VL-8B (Bai et al., 2025a) with different reward functions on COCO-LN500 (Pont-Tuset et al., 2020) and DOCCI500 (Onoe et al., 2024). Re^{3}Cap consistently improves over GRPO across most metrics, especially on the Relations QA task.

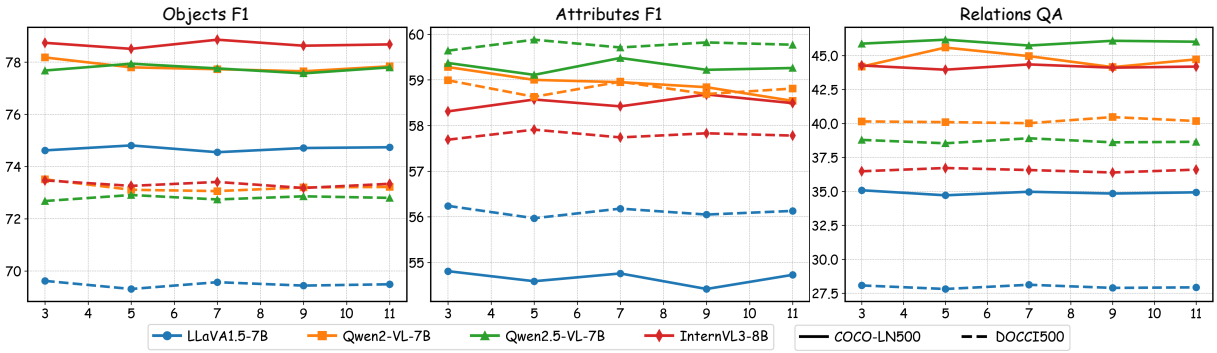

Figure 5: **Robustness study of the retrieval number K.** We evaluate Re^{3}Cap with different retrieval numbers K on COCO-LN500 (Pont-Tuset et al., 2020) and DOCCI500 (Onoe et al., 2024) across multiple LVLMs. The performance remains stable across different values of K, demonstrating that our method is robust to the choice of retrieval number.

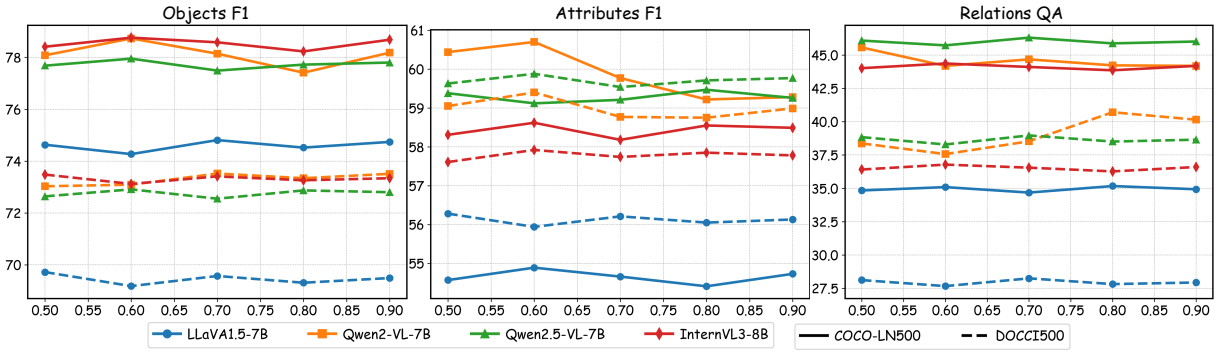

Figure 6: **Robustness study of the similarity threshold τ.** We evaluate Re^{3}Cap with different similarity thresholds τ for graph edge construction on COCO-LN500 (Pont-Tuset et al., 2020) and DOCCI500 (Onoe et al., 2024) across multiple LVLMs. The results show only minor variations across different thresholds, indicating that our method is robust to the choice of graph construction threshold.

Image Encoder	Text Encoder	Objects			Attributes			Relations
		Precision	Recall	F1	Precision	Recall	F1	QA
DINOv3 ViT-L/16	MiniLM	79.24	77.23	77.77	70.48	55.89	59.03	45.13
	MPNet	80.60	76.88	78.23	71.34	54.82	58.57	44.68
OpenCLIP ViT-H/14	MiniLM	79.77	77.03	77.91	69.99	55.29	58.49	45.04
	MPNet	79.62	77.70	78.18	71.00	56.03	59.28	44.19

Table 5: **Robustness study of diverse encoders on COCO-LN500 (Pont-Tuset et al., 2020) using Qwen2-VL-7B (Wang et al., 2024b) with CIM (Jia et al., 2026) as the reward function.** Our method maintains stable performance in image captioning when leveraging various encoders as the retrieval model. The experimental results indicate that our proposed Re^{3}Cap is robust to diverse encoders.