



## RAGCap: retrieval-augmented generation for style-aware remote sensing image captioning without fine-tuning

Yakoub Bazi, Mohamad M. Al Rahhal & Mansour Zuair

**To cite this article:** Yakoub Bazi, Mohamad M. Al Rahhal & Mansour Zuair (2025) RAGCap: retrieval-augmented generation for style-aware remote sensing image captioning without fine-tuning, International Journal of Remote Sensing, 46:22, 8903-8918, DOI: [10.1080/01431161.2025.2575514](https://doi.org/10.1080/01431161.2025.2575514)

**To link to this article:** <https://doi.org/10.1080/01431161.2025.2575514>



Published online: 28 Oct 2025.



Submit your article to this journal [↗](#)



Article views: 28



View related articles [↗](#)



View Crossmark data [↗](#)



# RAGCap: retrieval-augmented generation for style-aware remote sensing image captioning without fine-tuning

Yakoub Bazi<sup>a</sup>, Mohamad M. Al Rahhal<sup>b</sup> and Mansour Zuair<sup>a</sup>

<sup>a</sup>Computer Engineering Department, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia; <sup>b</sup>Applied Computer Science Department, College of Applied Computer Science, King Saud University, Riyadh, Saudi Arabia

## ABSTRACT

Recently, generalist vision-language models (VLMs) have proven to be highly versatile across various tasks, making them indispensable in computer vision and natural language processing. Their broad pre-training enables robust performance across many domains. When it comes to tailoring these models for specific downstream tasks such as remote sensing (RS) image captioning, the traditional approach has been fine-tuning. Fine-tuning, however, is often computationally expensive and may compromise the model inherent generalization capabilities by over-specializing on limited domain-specific datasets. This paper investigates Retrieval-Augmented Generation (RAG) as an alternative strategy for adapting generalist VLMs to RS captioning without requiring fine-tuning. We introduce RAGCap, a retrieval-augmented framework that leverages similarity-based retrieval to select relevant image-caption pairs from the training dataset. These examples are then combined with the target image within a carefully designed prompt structure, guiding the generalist VLM to generate stylistically coherent RS captions to the training dataset. While our implementation utilizes SigLIP for retrieval and Qwen2VL as the base VLM, the proposed framework is universal and applicable to other models. Extensive evaluations on four RS benchmark datasets reveal that RAGCap achieves competitive performance compared to traditional fine-tuning approaches. Our findings suggest RAG methods like RAGCap offer a scalable, practical alternative to fine-tuning for domain adaptation in RS image captioning. Code will be available at: <https://github.com/BigData-KSU/RAGCap>.

## ARTICLE HISTORY

Received 17 June 2025

Accepted 11 October 2025

## KEYWORD

Captioning; multi-image VLMs; remote sensing; RAG

## 1. Introduction

Image captioning is an important task in remote sensing (RS) as it enables the automated generation of textual descriptions for aerial imagery. This facilitates applications such as environmental monitoring, urban planning, and disaster management. Traditional methods for this task often rely on specialized models specifically designed for RS imagery (Bashmal et al. 2023). These models, while effective, need to be retrained for each dataset, making them less versatile for real-world deployment where applications are diverse.

Recently, the advent of large vision-language models (VLMs) has opened new possibilities for image captioning by leveraging their ability to generalize across diverse vision-language tasks (H. Liu et al. 2023). These models trained on massive multimodal datasets of images and texts capture rich cross-modal representations, enabling them to handle a variety of tasks including classification, question answering, and caption generation within a single unified framework. Compared to traditional approaches, large VLMs offer enhanced adaptability and robust performance in real-world contexts. This makes them particularly appealing for RS scenarios where images can vary significantly in content, scale, and resolution.

Adapting generalist VLMs to specialized downstream tasks typically begins with domain-specific fine-tuning, where models are further trained on tailored datasets. However, these datasets are often limited in size, and the fine-tuning process can lead to several challenges. Specifically, fine-tuning on small datasets may incur high computational costs and significantly increase the risk of overfitting, causing the model to adapt too closely to the training data at the expense of broader generalization. To mitigate these limitations, retrieval-augmented generation (RAG) methods have emerged as an effective alternative. Initially, successful within natural language processing tasks (Zhou and Long 2023), RAG methods work by retrieving relevant context and incorporating it into the input, thereby enhancing performance without the need for expensive fine-tuning. By leveraging an external repository of knowledge, RAG not only improves model performance but also preserves the inherent generalization capabilities that are often compromised during fine-tuning. This approach reduces computational overhead and overfitting risks, while maintaining the model broad adaptability for a wide range of downstream tasks.

This concept has been recently extended to multimodal domains, particularly in vision tasks. For example, the authors in (Zhou and Long 2023) introduces a retrieval mechanism that leverages region-of-interest, and triplet-based visual features to guide the captioning process to bridge the gap between semantic content and linguistic expression. In (J. Li et al. 2024), the authors enhance open-world captioning by retrieving object names from an external visual memory and prompting large language models to incorporate these terms into descriptions, thus enabling more accurate description of unfamiliar objects. Furthermore, the authors in (Lyu et al. 2025) address the challenge of fine-grained and realistic image generation by introducing a self-reflective contrastive learning framework. It retrieves visually aligned samples to refine generation, thereby improving object fidelity and reducing hallucinations in synthesized images. For a more comprehensive discussion of these advancements, the readers can refer to (X. Zheng et al. 2025).

Despite these advances in the general computer vision community, RAG remains largely unexplored for RS imagery, whose unique characteristics pose distinct challenges and opportunities for retrieval-based methods. Motivated by the success of multi-image VLMs (e.g. BLIP2 (J. Li et al. 2023), and Deepseek-VL2 (Wu et al. 2024)) in processing and reasoning about scenes across multiple images, we propose RAGCap, a novel, fine-tuning-free approach to RS image captioning. RAGCap leverages the Qwen2VL (P. Wang et al. 2024) architecture for its robust multi-image processing capabilities, injecting contextual information at inference time. Specifically, RAGCap retrieves the most similar images and corresponding captions from a training set using a retrieval model (e.g. SigLIP (Zhai et al. 2023)) and provides them along with the target RS image to Qwen2VL (P. Wang et al. 2024). A carefully crafted prompt then guides the model to produce a caption that reflects the style of the retrieved examples while highlighting the unique features of the test

image. As a result, RAGCap generates contextually aligned captions without computationally expensive model updates, preserving the versatility of the pre-trained VLM. Experimental results on three benchmark datasets confirm the competitiveness of this solution compared to both specialist captioning models and generalist models fine-tuned on various downstream tasks.

The main contributions of this paper are as follows:

- We introduce RAGCap, a retrieval-augmented generation framework for RS image captioning that avoids computationally expensive fine-tuning.
- We demonstrate that incorporating contextually relevant retrievals significantly enhances caption quality, allowing generalist VLMs to achieve performance competitive with specialized fine-tuned models.
- Through extensive experiments conducted on four benchmark RS datasets, we validate the effectiveness and versatility of the RAGCap framework.

The remainder of the paper is structured as follows. [Section II](#) describes in detail the proposed framework. [Section III](#) presents extensive experimental evaluations on four benchmark RS captioning datasets, analysing performance metrics and conducting ablation studies. Finally, [Section V](#) provides concluding remarks and discusses potential future research directions.

## 2. RagCap framework

Given a training set  $D = \{(I_i, C_i)\}_{i=1}^N$  consisting of  $I_i$  image-caption pairs, where  $I_i$  represents an aerial image and  $C_i = \{c_{ij}\}_{j=1}^m$  denotes its associated set of  $m$  multiple captions. Unlike existing solutions that require task-specific fine-tuning, our approach leverages pre-trained VLMs in a RAG framework. This framework uses dense visual embedding matching, given a query image  $I_q$  to retrieve semantically similar images and their captions from the training set, followed by context-aware language generation to produce a caption  $C_q$  that captures the content of the query image while reflecting the domain-specific style of the retrieved examples.

### 2.1. Visual embedding and retrieval using SigLIP

RAGCap utilizes a retriever model to identify semantically similar context images to the query image from the training set. While the architecture is model-independent and can integrate various retriever models, we employ SigLIP (google/siglip-so400m-patch14-384), a vision-language pre-training (VLM) model with a dual transformer architecture comprising a vision encoder and a language encoder. Pre-trained on 400 M image-text pairs, SigLIP is specifically designed for image-text alignment, leveraging a pairwise sigmoid loss that directly optimizes pairwise similarities, making it particularly effective for retrieval tasks. The vision encoder of SigLIP processes images resized to  $384 \times 384$  pixels by dividing them into patches of size  $14 \times 14$  pixels. These patches are then passed through the encoder to produce  $d = 1152$  dimensional, semantically rich feature embedding.

For a query image  $I_q$ , we compute its embedding feature as:

$$e_q = \text{normalize}(f(I_q)) \quad (1)$$

where  $f(\cdot)$  refers to the embedding function and  $\text{normalize}$  refers to L2 normalization, which scales the embedding vector. Then, we use the cosine similarity defined as:

$$\text{sim}(I_q, I_i) = e_q \cdot e_i = \sum_{j=1}^d e_{qj} \cdot e_{ij} \quad (2)$$

to measure similarity between  $I_q$  and all training images  $\{I_i\}_{i=1}^N$ . Then, we select the top  $k$  most similar images from the training set  $D$  with their captions by ranking the similarities  $\text{sim}(I_q, I_i)$ :

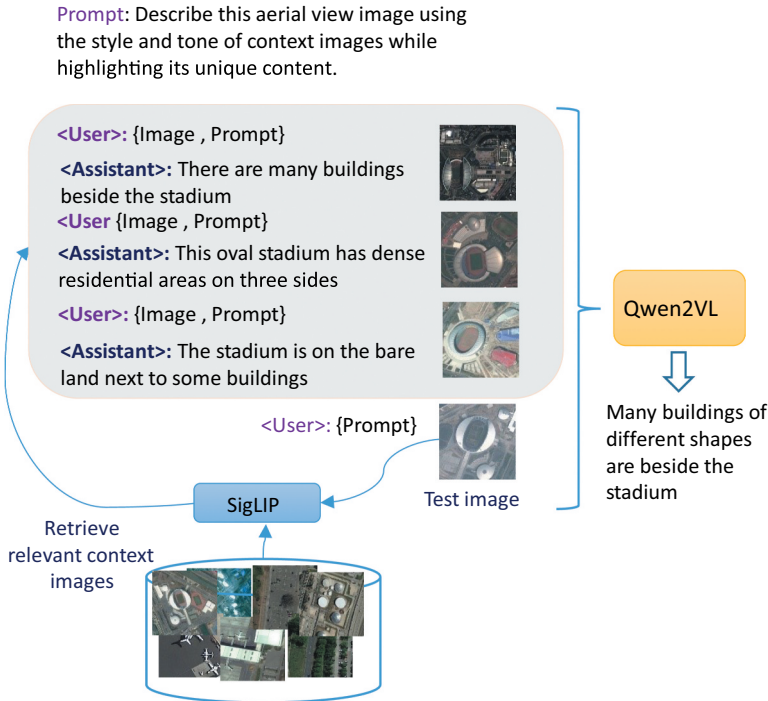
$$R(I_q, D) = \{(I_i, C_i)\}_{i=1}^k = \text{topk}\left(\{(I_i, C_i)\}_{i=1}^N\right) \quad (3)$$

It is worth mentioning that to efficiently the search and retrieval process, we utilize FAISS (Facebook AI Similarity Search) with a Flat Index. This index stores all embeddings in their full dimensionality and performs exact similarity searches. The Flat Index computes cosine similarities between the query embedding and all stored embeddings and ranks them, and retrieves the top most relevant image-caption pairs.

## 2.2. Context-aware caption generation using Qwen2VL

For context-aware caption generation, we use the Qwen2VL model, which integrates a vision transformer with approximately 675 M parameters. For language processing, Qwen2VL employs the powerful Qwen2 series of large language models (LLMs). In our main experiments, we adopt the Qwen2VL-7B with 7B parameters. To enhance the ability of the model to effectively perceive and comprehend visual information, the model comprises several innovations. These include the naive dynamic resolution mechanism, which enables the model to dynamically process images and videos of varying resolutions by converting them into an efficient sequence of visual tokens. Additionally, the model incorporates Multimodal Rotary Position Embedding (M-RoPE), which aligns positional information across text, images, and videos to ensure seamless multimodal integration. Supplementing these innovations is the ability of the model to handle multiple images simultaneously, making it exceptionally well-suited for context-aware generation tasks.

The caption generation process begins by retrieving the  $\text{topk}$  most similar images from the training set using the SigLIP vision encoder. These retrieved images, along with their associated captions, are combined with the query image and processed by Qwen2VL using a carefully structured prompt as shown in [Figure 1](#). Without requiring any fine-tuning, this approach allows Qwen2VL to adaptively incorporate the style, semantics, and context of the retrieved examples into the generated caption. Each retrieved image provides additional context, to guide Qwen2VL in producing captions that are both contextually enriched and reflective of domain-specific styles. The probability of generating the caption  $C_q = \{c_1, c_2, \dots, c_T\}$  of length  $T$  for the query image  $I_q$  is expressed as:



**Figure 1.** Overview of RAGCap: SigLIP retrieves similar images with captions, and Qwen2VL generates a caption for the test image, blending the style of the context captions while emphasizing its unique content.

$$P\left(C_q|I_q, \{(I_i, C_i)\}_{i=1}^k\right) = \prod_{t=1}^T P\left(c_t|c_{<t}, I_q, \{(I_i, C_i)\}_{i=1}^k\right) \quad (4)$$

where  $T$  denotes the total number of tokens in the caption. In the following, we outline the key algorithmic steps of RAGCap. In our implementation, we adopt SigLIP as the retriever and Qwen2VL-7B, with its inherent multi-image input capability, as the context-aware captioner. The architecture of the framework is modular and independent, allowing the use of other models with similar capabilities. This makes the architecture universal and adaptable to a variety of retrieval and captioning tasks.

---

**Algorithm: RAGCap**

---

**Inputs:**

- Query image:  $I_q$
- Training dataset of images and captions  $D = \{(I_i, C_i)\}_{i=1}^N$
- Pre-trained vision encoder: SigLIP
- Vision language model: Qwen2VL-7B
- Number of retrieved context examples:  $k$
- Prompt:  $P$

**Output:**

- Generated caption  $C_q$  for the query image  $I_q$

**Step 1: Retrieve context images using SigLIP**

1. Extract visual embedding  $e_q$  for  $I_q$  and all training images  $\{e_i\}_{i=1}^N$
- 

(Continued)

**Algorithm: RAGCap**

2. Compute cosine similarity scores  $\text{sim}(l_q, l_i) = e_q, e_i$  between  $l_q$  and  $\{l_i\}_{i=1}^N$
3. Retrieve similar images and their captions  $\text{topk}(\{(l_i, C_i)\}_{i=1}^N)$

**Step 2: Caption aware generation using Qwen2VL-7B**

1. Define task-specific prompt to guide the model:  
 $P = \text{"Describe this aerial view image using the style and tone of context image while highlighting its unique content"}$
2. Encode the query image and context images with their captions into an adequate prompt chat template:  
 $\{(P, l_1, c_{11}, c_{12}, \dots, c_{1m}), \dots, (P, l_k, c_{k1}, c_{k2}, \dots, c_{km})\}, (P, l_q)$
3. Pass the constructed prompt template to Qwen2VL-7B to generate the caption  $C_q$ .

### 3. Experiments results

#### 3.1. Dataset description and experiment setup

To evaluate RAGCap, we utilize four benchmark RS datasets (Cheng et al. 2022): Sydney, UCM, CapRS, and NWPU datasets as shown in Figure 2. Each image in these datasets is paired with five manually annotated captions. The Sydney dataset consists of 613 images, divided into 497 for training, and 58 for each testing validation; however, only the training and test splits are utilized in our experiments. Likewise, the UCM dataset includes 1,680



**Figure 2.** Example of RS images with captions from: (a) Sydney, (b) UCM, (c) CapRS, and (d) NWPU datasets.



training images, 210 for testing and 210 for validation. The CapRS dataset is introduced in (Abdullah et al. 2020) for text-image retrieval. It consists of 2,144 images sampled from different scene classification datasets. Each image was annotated with 5 human-written captions, totalling 10,720 diverse descriptions. Images were uniformly sampled (16 per class) to ensure class balance across 134 scene categories. Captions focus on dominant objects, avoid unnecessary details (e.g. colour, exact counts), and follow concise structural rules. In our experiments, we use 1072 images for training and 1072 images for testing. The NWPU dataset, the largest among them, contains 31,500 images, split into 25,200 for training, 3,150 for testing, and 3,150 for validation with images of resolutions ranging from 0.2 to 30 metres.

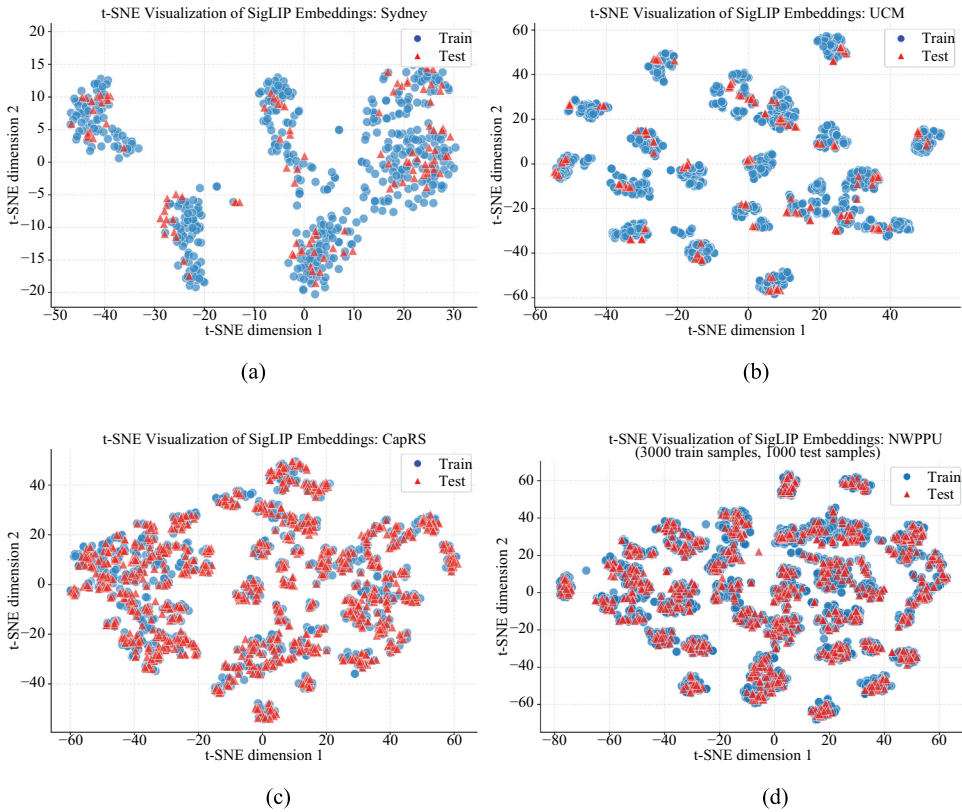
We evaluate the RAGCap using standard captioning metrics, including BLEU (Bilingual Evaluation Understudy), ROUGE (Recall-Oriented Understudy for Gisting Evaluation), CIDEr (Consensus-based Image Description Evaluation), and SPICE (Semantic Propositional Image Captioning Evaluation). BLEU measures the precision of  $n$ -grams in the generated captions compared to the reference captions, focusing on word overlap. ROUGE emphasizes recall by evaluating phrase,  $n$ -gram, and sequence overlaps between generated and reference captions. CIDEr assesses the semantic relevance of captions using TF-IDF (Term Frequency-Inverse Document Frequency) weighted  $n$ -grams, rewarding captions that are both accurate and diverse. Finally, SPICE evaluates the semantic structure by comparing objects, attributes, and relationships in the captions, offering a deeper analysis of caption meaning beyond simple word matches.

### **3.2. Evaluation of Qwen2VL zero-shot and RAGCap under varying context images $k$**

In our RAGCap framework, SigLIP is central to the retrieval process, generating high-dimensional embeddings that capture the semantic and visual features of remote sensing images. To qualitatively assess these embeddings, we applied t-SNE, a dimensionality reduction technique, to project the feature space onto a two-dimensional map. The resulting visualization depicted in Figure 3 reveals distinct clusters for each dataset, highlighting their unique scene characteristics. For instance, the UCM dataset displays well-separated clusters corresponding to its diverse land-use and land-cover classes, whereas the Sydney dataset forms a relatively dense and compact cluster, indicative of its more homogeneous urban features. The NWPU dataset exhibits a moderately scattered distribution of clusters, reflecting its broader variety of land-cover types and higher scene complexity, which SigLIP effectively differentiates despite overlapping categories. Finally, the CapRS dataset shows a less clear clustering pattern, characterized by multiple, sparsely populated clusters due to each scene type being represented only a few times, leading to a more dispersed embedding distribution. These observations confirm that SigLIP exhibits an interesting capacity to capture meaningful similarities among images, even though it is not specifically trained on RS imagery. This ability enables the retrieval of contextually relevant image-caption pairs, which is fundamental to RAGCap.

Table 1 presents the performances on the Sydney, UCM, CapRS, and NWPU datasets, evaluated using BLEU, METEOR, ROUGE, CIDEr, and SPICE metrics (all reported in





**Figure 3.** t-SNE embedding of SigLIP for: (a) Sydney, (b) UCM, (c) NWPU and (d) CapRS datasets.

percentages). In the zero-shot scenario ( $k=0$ ), the base model (Qwen2VL-7B) exhibits modest performance. For example, on the Sydney dataset, it yields a BLEU-1 score of 27.02, a CIDEr score of 6.50, and METEOR, ROUGE, and SPICE scores of 15.81, 27.17, and 11.89, respectively. Similarly, the Merced dataset shows a BLEU-1 score of 25.68 and a CIDEr score of 15.36 with METEOR, ROUGE, and SPICE scores of 16.65, 26.15, and 12.07, respectively. For CapRS, it shows a BLEU-1 score of 29.38 and a CIDEr score of 6.70 with METEOR, ROUGE, and SPICE scores of 17.28, 25.47, and 12.93, respectively. For NWPU it achieves corresponding scores of 38.10 (BLEU-1), 15.52 (CIDEr), 15.77 (METEOR), 27.96 (ROUGE), and 10.98 (SPICE). These zero-shot results serve as a baseline, reflecting the performance of a pre-trained vision-language model without domain-specific adaptation.

In contrast, RAGCap employs a RAG paradigm that dynamically retrieves contextual image-caption pairs from the training set to guide caption generation without fine-tuning. By varying the context size ( $k$ ), we observe substantial performance improvements across all datasets, with peak performance occurring in general at  $k=9$ . On the Sydney dataset, RAGCap achieves a BLEU-1 score of 80.52, CIDEr of 239.32, and METEOR, ROUGE, and SPICE scores of 41.66, 73.14, and 47.42, respectively. Similarly, on the UCM dataset, RAGCap attains 88.75 BLEU-1, 367.08 CIDEr, and METEOR, ROUGE, and SPICE scores of 48.05, 83.86, and 53.14. For CapRS, it achieves 66.69 BLEU-1, 118.53 CIDEr, and 25.64,

**Table 1.** Performance of RaGCap in terms of BLEU, METOR, ROUGE, cider, and SPICE scores (%) on (a) Sydney, (b) UCM, (c) CapRS, and (d) NWPU datasets. The table also reports the impact of the retriever and generator on model performance, along with comparisons against fine-tuning Qw2nVL using LoRA Method.

(a)									
Method	Context	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METOR	ROUGE	CIDEr	SPICE
Zero-shot (Qwen2vl-7B+SigLIP)	$k=0$	27.02	14.44	08.74	04.08	15.81	27.17	06.50	11.89
	$k=1$	73.03	6763	55.74	48.73	35.98	64.63	193.67	38.84
	$k=5$	79.30	72.48	65.99	59.46	39.77	71.00	224.55	46.93
	$k=9$	80.52	73.53	67.37	61.33	41.66	73.14	239.32	47.42
	$k=15$	80.08	72.78	65.91	69.37	40.73	72.62	225.78	46.98
	$k=20$	80.36	73.78	67.31	60.75	41.89	73.75	237.55	47.84
FineTune Qwen2vl-7B (LoRA)		79.91	71.96	64.16	56.97	41.50	73.13	241.51	48.01
Qwen2vl-7B+Random	$k=9$	67.15	56.75	49.90	44.46	31.85	58.70	117.24	34.67
Qwen2vl-7B+Git-RSCLIP	$k=5$	81.78	75.34	69.28	63.31	41.94	73.31	253.57	48.43
	$k=9$	80.09	73.55	67.58	61.93	42.07	72.97	245.43	47.28
Qwen2vl-2B+SigLIP	$k=9$	78.22	70.07	62.79	56.32	40.98	71.73	232.14	45.19
(b)									
Method	Context	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METOR	ROUGE	CIDEr	SPICE
Zero-shot (Qwen2vl-7B+SigLIP)	$k=0$	25.68	13.81	7.78	4.20	16.65	26.15	15.36	12.07
	$k=1$	82.73	74.61	67.55	61.05	40.01	73.27	284.64	43.79
	$k=5$	87.65	81.64	76.21	70.99	46.89	82.65	359.57	52.69
	$k=9$	88.75	82.21	76.25	70.65	48.05	83.86	367.08	53.14
	$k=15$	87.62	81.91	76.57	71.59	47.77	82.92	365.89	52.49
	$k=20$	88.01	81.70	75.93	70.35	47.32	83.24	358.78	53.28
FineTune Qwen2vl-7B (LoRA)		87.72	81.39	75.59	70.38	47.05	83.16	361.44	51.20
Qwen2vl-7B+Random	$k=9$	62.92	50.40	40.78	33.28	26.30	52.47	109.22	28.11
Qwen2vl-7B+ Git-RSCLIP	$k=5$	88.34	82.22	76.34	70.37	46.43	82.42	356.47	53.60
	$k=9$	87.67	81.47	75.72	70.20	47.05	82.31	355.16	53.03
Qwen2vl-2B+SigLIP	$k=9$	83.84	76.68	70.33	64.54	79.72	343.53	343.57	51.73
(c)									
Method	Context	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METOR	ROUGE	CIDEr	SPICE
Zero-shot (Qwen2vl-7B+SigLIP)	$k=0$	29.38	12.51	6.27	3.00	17.28	25.47	6.70	12.93
	$k=1$	62.60	43.29	29.96	20.58	23.30	48.64	103.54	18.31
	$k=5$	66.70	47.23	33.09	22.74	25.31	51.16	117.32	19.66
	$k=9$	66.96	48.41	33.98	23.47	25.64	52.52	118.53	19.89
	$k=15$	66.83	48.47	34.37	24.06	25.42	52.91	117.64	19.82
	$k=20$	66.39	48.19	34.69	24.09	25.47	52.97	118.20	19.87
FineTune Qwen2vl-7B (LoRA)		65.84	47.85	33.92	24.02	25.64	52.86	119.50	19.20
Qwen2vl-7B+Random	$k=9$	54.81	32.54	19.84	12.03	18.03	40.47	56.49	13.67
Qwen2vl-7B+ Git-RSCLIP	$k=5$	67.89	49.39	35.18	24.67	26.67	53.14	126.18	20.89
	$k=9$	67.73	49.50	35.47	25.18	26.25	53.40	126.47	20.88
Qwen2vl-2B+SigLIP	$k=9$	67.03	48.83	34.65	24.42	25.75	52.87	119.50	19.76
(d)									
Method	Context	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METOR	ROUGE	CIDEr	SPICE
Zero-shot (Qwen2vl-7B+SigLIP)	$k=0$	38.10	17.32	7.99	3.62	15.77	27.96	15.52	10.98
	$k=1$	81.15	68.64	58.11	49.59	34.47	64.94	144.23	25.29
	$k=5$	86.24	76.19	67.21	59.45	39.84	72.96	176.96	29.24
	$k=9$	86.56	76.99	68.45	61.13	40.60	74.08	179.48	29.60
	$k=15$	86.63	77.23	68.84	61.62	41.34	74.59	182.37	29.81
	$k=20$	86.44	76.94	68.54	61.46	41.32	74.36	181.42	29.55
FineTune Qwen2vl-7B (LoRA)		88.42	79.78	72.10	65.43	43.32	77.15	195.70	31.20
Qwen2vl-7B+Random	$k=9$	72.07	52.04	36.18	23.78	22.01	49.07	24.11	14.90
Qwen2vl-7B+ Git-RSCLIP	$k=5$	87.78	78.51	70.07	62.80	41.63	76.10	183.73	30.59
	$k=9$	88.54	79.80	71.92	65.10	42.76	76.44	189.19	30.86
Qwen2vl-2B+SigLIP	$k=9$	84.26	74.09	65.42	58.05	39.51	71.56	170.73	28.85

52.52, and 19.87 in METEOR, ROUGE, and SPICE, respectively. On the other side on NWPU, it achieves 86.56 BLEU-1, 179.48 CIDEr, and 40.60, 74.08, and 29.60 in METEOR, ROUGE, and SPICE, respectively. Notably, increasing  $k$  beyond 9 yields may lead to marginal performance gains while incurring increased memory overhead. These results demonstrate that retrieving relevant context significantly enhances captioning quality, making RAGCap an efficient alternative to fine-tuned models for RS image captioning, while offering a balance between retrieval and generation without computationally expensive updates to model parameters.

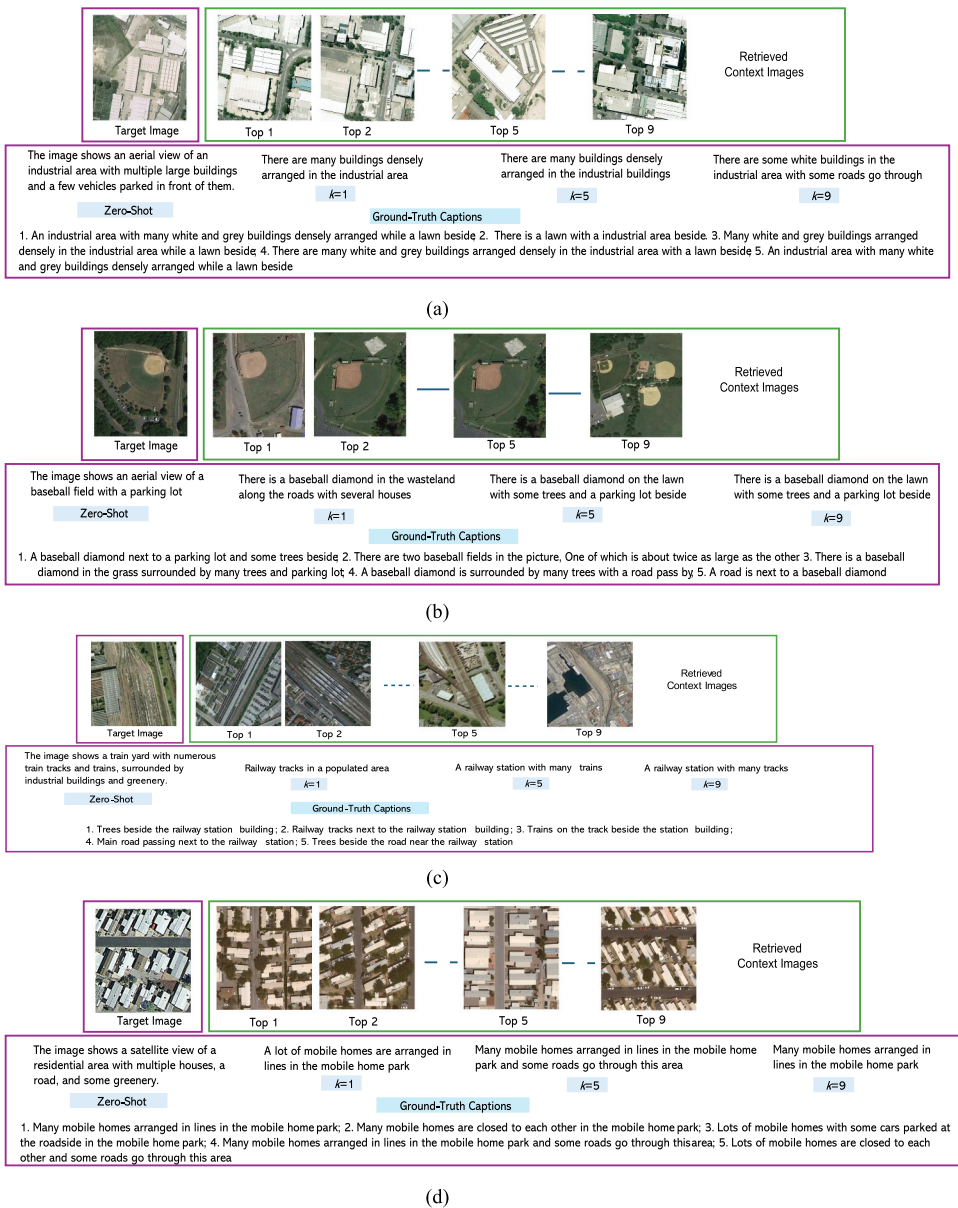
To demonstrate the effectiveness of the RAG paradigm with respect to fine-tuning, we compare RAGCap ( $k = 9$ ) against LoRA (E. J. Hu et al. 2022) fine-tuning of Qwen2VL-7B. For LoRA, we adopt a matrix rank = 16, scaling factor alpha = 32, and train for 5 epochs using the LLaMA Factory platform (Y. Zheng et al. 2024) across all four datasets. RAGCap achieves competitive results, outperforming fine-tuning on three out of four datasets: Sydney (BLEU-1: 80.52 vs 79.91), UCM with notable improvements (BLEU-1: 88.75 vs 87.72, CIDEr: 367.08 vs 361.44), and CapRS (BLEU-1: 66.96 vs 65.84). Only on the largest NWPU dataset where fine-tuning shows better performance (BLEU-1: 88.42 vs 86.56). These results demonstrate that RAGCap offers an efficient alternative to fine-tuning in particular for small downstream datasets, delivering competitive performances while eliminating the computational overhead of parameter updates and model finetuning.

For further analysis, we present in Figure 4 the qualitative captioning results generated by RAGCap with varying context image sizes on the Sydney, UCM and NWPU datasets. Initially, in a zero-shot setting without retrieval, Qwen2VL produces captions directly, often lacking domain-specific coherence. By incorporating context images, the prompt effectively guides Qwen2VL to generate captions that align with the stylistic and semantic patterns of the retrieved examples. This approach enhances descriptive accuracy and contextual relevance, as evident in comparison with the baseline results. This qualitative analysis highlights the advantage of retrieval-augmented captioning, demonstrating more enhanced and contextually adapted descriptions to RS imagery.

### 3.3. Impact of the retriever and context aware generator

To assess the sensitivity of RAGCap to both the retriever and the context-aware generator, we conducted two key experiments. First, we evaluate the impact of retrieval by replacing the SigLIP-based retriever with a random selection of training set images as context. The results demonstrate an improvement over the zero-shot baseline but remain significantly lower than those achieved with SigLIP, confirming the critical role of retrieval in obtaining contextually relevant images for guiding the caption generation process. For example, for  $k=9$  the BLEU-1 scores across datasets were markedly lower with random retrieval, dropping from 80.52 to 67.15 on Sydney, from 88.75 to 62.92 on UCM, from 99.96 to 54.81 on CapRS and from 86.56 to 72.07 on NWPU. Similarly, the same observation holds for other metrics showing substantial declines, reinforcing the necessity of an effective retriever.

For further analysis, we also evaluated another retriever adapted to the RS domain. In particular, we considered the Git-RSCLIP retriever introduced in the Text2Earth work (C. Liu et al. 2025). Git-RSCLIP is pretrained on the large-scale Git-10 M dataset (around 10.5 M RS image – text pairs). Its contrastive training objective aligns textual descriptions with RS



**Figure 4.** Qualitative captioning results produced by RAGCap using different number of context images on (a) Sydney, (b) UCM, (c) CapRS and (d) NWPU datasets.

images covering diverse spatial patterns and geographic regions. When comparing Git-RSCLIP against SigLIP, we observed consistent gains at both  $k=5$  and  $k=9$  across the different datasets. On the Sydney dataset, it achieved a CIDEr score of 253.57 at  $k=5$ , significantly higher than SigLIP, which yielded 224.55, while at  $k=9$  it yielded 245.43 versus 239.32 for SigLIP. For UCM, SigLIP showed slightly better CIDEr results, for example, at  $k=9$  it yielded 367.08 versus 355.16 for Git-RSCLIP. On CapRS, Git-RSCLIP clearly outperformed SigLIP at both  $k=5$  (126.18 versus 117.32) and  $k=9$  (126.47 versus

118.53). Finally, on NWPU, Git-RSCLIP showed better CIDEr at  $k = 5$  (183.73 versus 176.96) and  $k = 9$  (189.19 versus 179.48), with better BLEU and SPICE scores.

In summary, the experiments show that replacing the general-purpose retriever SigLIP with RS-specific retrievers like Git-RSCLIP can yield better results. Indeed, these models, pretrained on large-scale RS imagery and text pairs, are better at capturing the spatial and semantic structures unique to RS. As a result, they can provide more relevant retrievals, which directly translate into improved style captioning. This clearly confirms that the proposed RAGCap with RS adapted retrievers can be more effective than general-purpose ones.

In the third experiment, we examine the impact of the caption generator by replacing Qwen2VL-7B with the lighter Qwen2VL-2B model while maintaining SigLIP as the retriever. Across datasets, the results were lower compared to Qwen2VL-7B, highlighting the importance of a strong generative model for achieving high-quality captions. For instance, at  $k = 9$ , scores with Qwen2VL-2B were consistently lower than those with Qwen2VL-7B except for the CapRS dataset. These findings confirm that both retrieval and generation components play crucial roles in the effectiveness of RAGCap. A high-quality retriever ensures relevant context, while a capable generator fully leverages this context to produce accurate and stylistically aligned captions.

3.4. Failure case analysis

To understand better the limitation of RAGCap, we show in Figure 5 some representative failure cases from the Sydney and NWPU datasets. In Sydney, the most evident errors include object hallucination ('solar panels' versus a parking lot), salient-structure omission (river described without the highway), and domain-term intrusion ('runways' versus a meadow with roads). In NWPU, the failures highlight object hallucination (clouds foregrounded while farmlands are ignored), salience omission (a freeway simplified to 'empty motorway'), and domain-term intrusion ('palace' versus a church complex).



Figure 5. Representative failure cases form: Sydney dataset (first row) and NWPU dataset (second row).



landmark misclassification (a church complex mislabelled as a palace). These cases demonstrate that the most critical weaknesses stem from hallucinating non-existent objects, omitting key spatial relations, and misapplying domain-specific semantics.

### 3.5. Comparisons to SOTA methods

We compare RAGCap with both specialist RS-captioning models and generalist VLMs reported in the literature and mainly based on model finetuning on the downstream task as shown in Table 2. Specialist models, such as MLCA-NET (Cheng

**Table 2.** Comparison between RAGCap and state-of-the-art methods based on finetuning on: (A) Sydney, (B) UCM, and (C) NWPU datasets.

(A)								
Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METOR	ROUGE	CIDEr	SPICE
MLCA-NET (Cheng et al. 2022)	83.10	74.20	65.90	58.00	39.00	71.10	232.40	40.90
Post processing (Hoxha, Scuccato, and Melgani 2023)	78.37	69.85	63.22	57.17	39.49	71.06	255.53	—
RS-CapRet (Silva et al. 2024)	78.70	70.00	62.80	56.40	38.80	70.70	239.20	43.40
SD-RSIC (Sumbul, Nayak, and Demir 2021)	72.4	62.1	53.2	45.1	34.2	63.6	139.5	—
SVM-D BOW (Hoxha and Melgani 2022)	77.87	68.35	60.23	53.05	37.97	69.92	227.22	—
GVFGA+LSGA (Zhang et al. 2022)	76.81	68.46	61.45	55.04	38.66	70.3	245.22	—
SVM-D CONC (Hoxha and Melgani 2022)	75.47	67.11	59.7	53.08	36.43	67.46	222.22	—
RSGPT (Y. Hu et al. 2025)	82.26	75.28	68.57	62.23	41.37	74.77	273.08	—
HCNet (Yang et al. 2024)	76.86	71.09	65.73	61.02	39.8	71.72	247.14	—
RAG-Cap (SiGLIP)	80.52	73.53	67.37	61.33	41.66	73.14	239.32	47.42
RAG-Cap (Git-RSLIP)	80.09	73.55	67.58	61.93	42.07	72.97	245.43	47.28
(B)								
Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METOR	ROUGE	CIDEr	SPICE
MLCA-NET (Cheng et al. 2022)	82.60	77.00	71.70	66.80	43.50	77.20	324.00	47.30
Post processing (Hoxha, Scuccato, and Melgani 2023)	79.39	72.98	67.44	62.62	40.80	74.06	309.64	—
SD-RSIC (Sumbul, Nayak, and Demir 2021)	74.8	66.4	59.8	53.8	39	69.5	213.2	—
RTRMN (semantic) (B. Wang et al. 2020)	55.26	45.15	39.62	35.87	25.98	55.38	180.25	—
RTRMN (statistical) (B. Wang et al. 2020)	80.28	73.22	68.21	63.93	42.58	77.26	312.7	—
GVFGA+LSGA (Zhang et al. 2022)	83.19	76.57	71.03	65.96	44.36	78.45	332.7	—
SVM-D BOW (Hoxha and Melgani 2022)	76.35	66.64	58.69	51.95	36.54	68.01	271.42	—
SVM-D CONC (Hoxha and Melgani 2022)	76.53	69.47	64.17	59.42	37.02	68.77	292.28	—
RSGPT (Y. Hu et al. 2025)	86.12	79.14	72.31	65.74	42.21	78.34	333.23	—
RS-CapRet (Silva et al. 2024)	84.30	77.90	72.20	67.00	47.20	81.70	354.80	52.50
RAG-Cap (ours)	88.75	82.21	76.25	70.65	48.05	83.86	367.08	53.14
RAG-Cap (Git-RSLIP)	87.67	81.47	75.72	70.20	47.05	82.31	355.16	53.03
(C)								
Method	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METOR	ROUGE	CIDEr	SPICE
MLCA-NET (Cheng et al. 2022)	74.50	62.40	54.10	47.80	33.70	60.10	116.40	28.50
RS-CapRet (Silva et al. 2024)	87.10	78.70	71.70	65.60	43.60	77.60	192.90	31.10
MTT (Ren et al. 2022)	78.75	65.71	56.88	50.13	30.87	62.5	117.74	—
SALCap (Guo et al. 2025)	79.25	65.92	56.88	49.94	30.99	63.17	118.04	—
DeCap (W. Li et al. 2023)	68.80	50.50	37.70	28.3	23.9	48.90	63.60	—
HCNet (Yang et al. 2024)	78.49	64.90	56.48	52.06	30.53	64.09	131.06	—
ViECap (Fei et al. 2023)	56.73	38.81	28.17	21.62	20.37	42.08	49.86	—
Clipcap (Silva et al. 2024)	83.94	74.21	66.19	59.54	41.42	73.85	172.65	—
RAG-Cap (ours)	86.56	76.99	68.45	61.13	40.60	74.08	179.48	29.60
RAG-Cap (Git-RSLIP)	88.54	79.80	71.92	65.10	42.76	76.44	189.19	30.86

et al. 2022) and post-processing approaches (Hoxha, Scuccato, and Melgani 2023), are explicitly optimized for RS. Generalist models, such as RS-CapRet integrates image captioning and text-image retrieval in a unified framework. Unlike these models, RAGCap does not require fine-tuning; instead, it retrieves stylistically and semantically relevant examples at test time to enhance caption generation. Despite avoiding parameter updates, RAGCap achieves competitive or superior performance across BLEU, METEOR, ROUGE, CIDEr, and SPICE on the Sydney, UCM, and NWPU datasets. By integrating domain-specific information dynamically at inference, RAGCap enables efficient captioning without the computational overhead or risk of overfitting associated with model retraining.

It is worth noting that RAGCap does not require any training and runs fully at test time. So, it falls in the category of test-time methods. The retriever stays on the GPU and needs about 1 GB, while the FAISS index is stored on the CPU and is very small (tens of MBs). The main memory use comes from the generator Qwen2-VL-7B, which takes about 15–17 GB. Adding more context images increases the cache size; for example, for  $k=5$  memory is around 16–18 GB, and with  $k=9$  it is around 18–20 GB. Retrieval is almost instant ( $< 0.1$  s), so the total inference time is about 2 seconds per image on a 48 GB GPU.

## 4. Conclusions

In this work, we harness the power of multimodal VLMs for RS image captioning. Unlike existing solutions that require fine-tuning on domain-specific datasets, we propose RAGCap that adapts pre-trained VLMs to RS captioning without modifying model parameters. Our experiments on four benchmark datasets demonstrate that RAGCap achieves competitive performance relative to both specialist models and fine-tuned generalist models, while significantly reducing computational costs and avoiding overfitting risks. Future research avenues include exploring different retrievers and generative models with multi-image capabilities (both open- and closed-source), refining prompt designs for improved test-time performance, and extending RAG strategies to other RS tasks.

## Acknowledgement

This work was supported by the Ongoing Research Funding Program King Saud University, Riyadh, Saudi Arabia (ORF-2025-995).

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

This work was supported by the Ongoing Research Funding Program King Saud University, Riyadh, Saudi Arabia (ORF-2025-995).



## References

- Abdullah, T., Y. Bazi, M. M. Al Rahhal, M. L. Mekhalafi, L. Rangarajan, and M. Zuair. 2020. "Deep Bidirectional Triplet Network for Matching Text to Remote Sensing Images." *Remote Sensing* 12 (3, Art. no. 3). Jan. <https://doi.org/10.3390/rs12030405>.
- Bashmal, L., Y. Bazi, F. Melgani, M. M. Al Rahhal, and M. A. Al Zuair. 2023. "Language Integration in Remote Sensing: Tasks, Datasets, and Future Directions." *IEEE Geoscience and Remote Sensing Magazine* 11 (4): 63–93. Dec. <https://doi.org/10.1109/MGRS.2023.3316438>.
- Cheng, Q., H. Huang, Y. Xu, Y. Zhou, H. Li, and Z. Wang. 2022. "NWPU-Captions Dataset and MLCA-Net for Remote Sensing Image Captioning." *IEEE Transactions on Geoscience & Remote Sensing* 60:1–19. <https://doi.org/10.1109/TGRS.2022.3201474>.
- Fei, J., T. Wang, J. Zhang, Z. He, C. Wang, and F. Zheng. 2023. "Transferable Decoding with Visual Entities for Zero-Shot Image Captioning." Presented at the Proceedings of the IEEE/CVF International Conference on Computer Vision October 2023 Paris, France, 3136–3146. Accessed May 28, 2025. [https://openaccess.thecvf.com/content/ICCV2023/html/Fei\\_Transferable\\_Decoding\\_with\\_Visual\\_Entities\\_for\\_Zero-Shot\\_Image\\_Captioning\\_ICCV\\_2023\\_paper.html](https://openaccess.thecvf.com/content/ICCV2023/html/Fei_Transferable_Decoding_with_Visual_Entities_for_Zero-Shot_Image_Captioning_ICCV_2023_paper.html).
- Guo, Z., H. Liu, Z. Ren, L. Jiao, S. Gou, and R. Li. 2025. "Attribute-Based Learning for Remote Sensing Image Captioning in Unseen Scenes." *Remote Sensing* 17 (7, Art. no. 7). Jan. <https://doi.org/10.3390/rs17071237>.
- Hoxha, G., and F. Melgani. 2022. "A Novel SVM-Based Decoder for Remote Sensing Image Captioning." *IEEE Transactions on Geoscience & Remote Sensing* 60:1–14. <https://doi.org/10.1109/TGRS.2021.3105004>.
- Hoxha, G., G. Scuccato, and F. Melgani. 2023. "Improving Image Captioning Systems with Postprocessing Strategies." *IEEE Transactions on Geoscience & Remote Sensing* 61:1–13. <https://doi.org/10.1109/TGRS.2023.3281334>.
- Hu, E. J. Shen Y. Wallis P. Allen-Zhu Z. Li Y. Wang S. Chen W. 2022 "LORA: Low-Rank Adaptation of Large Language Models." Presented at the International Conference on Learning Representations 1(2) (ICLR) , Oct. Accessed May 31, 2025. [Online]. Available: <https://openreview.net/forum?id=nZeVKeeFYf9>.
- Hu, Y., J. Yuan, C. Wen, X. Lu, Y. Liu, and X. Li. 2025. "RSGPT: A Remote Sensing Vision Language Model and Benchmark." *ISPRS Journal of Photogrammetry & Remote Sensing* 224:272–286. Jun. <https://doi.org/10.1016/j.isprsjprs.2025.03.028>.
- Li, J., D. Li, S. Savarese, and S. Hoi. 2023. "BLIP-2: Bootstrapping Language-Image Pre-Training with Frozen Image Encoders and Large Language Models." Proceedings of the 40th International Conference on Machine Learning Honolulu, Hawaii, USA, In Proceedings of Machine Learning Research, Vol. 202. PMLR, edited by, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, 19730–19742. [Online]. Available: Jul. <https://proceedings.mlr.press/v202/li23q.html>.
- Li, J., D. M. Vo, A. Sugimoto, and H. Nakayama. 2024. "EVCap: Retrieval-Augmented Image Captioning with External Visual-Name Memory for Open-World Comprehension." Presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Seattle, WA, USA, 13733–13742. Accessed May 29, 2025. [Online]. Available: [https://openaccess.thecvf.com/content/CVPR2024/html/Li\\_EVCap\\_Retrieval-Augmented\\_Image\\_Captioning\\_with\\_External\\_Visual-Name\\_Memory\\_for\\_Open-World\\_CVPR\\_2024\\_paper.html](https://openaccess.thecvf.com/content/CVPR2024/html/Li_EVCap_Retrieval-Augmented_Image_Captioning_with_External_Visual-Name_Memory_for_Open-World_CVPR_2024_paper.html).
- Li, W., L. Zhu, L. Wen, and Y. Yang. 2023. Decap: Decoding CLIP Latents for Zero-Shot Captioning via Text-Only Training." Mar 06. <https://doi.org/10.48550/arXiv.2303.03032>.
- Liu, C., K. Chen, R. Zhao, Z. Zou, and Z. Shi. 2025. "Text2Earth: Unlocking Text-Driven Remote Sensing Image Generation with a Global-Scale Dataset and a Foundation Model." *IEEE Geoscience and Remote Sensing Magazine* 13 (3): 238–259. Sep. <https://doi.org/10.1109/MGRS.2025.3560455>.
- Liu, H., C. Li, Q. Wu, and Y. J. Lee. 2023. "Visual Instruction Tuning." *Advances in Neural Information Processing Systems* 36:1 34892–34916. Dec.
- Lyu, Y. Zheng X. Jiang L. Yan Y. Zou X. Zhou H. Hu X. 2025. RealRAG: Retrieval-Augmented Realistic Image Generation via Self-Reflective Contrastive Learning." *arXiv: arXiv:2502.00848*. May 12. <https://doi.org/10.48550/arXiv.2502.00848>.

- Ren, Z., S. Gou, Z. Guo, S. Mao, and R. Li. 2022. "A Mask-Guided Transformer Network with Topic Token for Remote Sensing Image Captioning." *Remote Sensing* 14 (12, Art. no. 12). Jan. <https://doi.org/10.3390/rs14122939>.
- Silva, J. D., J. Magalhães, D. Tuia, and B. Martins. 2024. "Large Language Models for Captioning and Retrieving Remote Sensing Images." arXiv: arXiv:2402.06475. <https://doi.org/10.48550/arXiv.2402.06475>.
- Sumbul, G., S. Nayak, and B. Demir. 2021. "SD-RSIC: Summarization-Driven Deep Remote Sensing Image Captioning." *IEEE Transactions on Geoscience & Remote Sensing* 59 (8): 6922–6934. Aug. <https://doi.org/10.1109/TGRS.2020.3031111>.
- Wang, B., X. Zheng, B. Qu, and X. Lu. 2020. "Retrieval Topic Recurrent Memory Network for Remote Sensing Image Captioning." *IEEE Journal of Selected Topics in Applied Earth Observations & Remote Sensing* 13:256–270. <https://doi.org/10.1109/JSTARS.2019.2959208>.
- Wang, P. 2024. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution." Oct 03. <https://doi.org/10.48550/arXiv.2409.12191>.
- Wu, Z. 2024. DeepSeek-VL2: Mixture-of-Experts Vision-Language Models for Advanced Multimodal Understanding." arXiv: arXiv:2412.10302. <https://doi.org/10.48550/arXiv.2412.10302>.
- Yang, Z., Q. Li, Y. Yuan, and Q. Wang. 2024. "HCNet: Hierarchical Feature Aggregation and Cross-Modal Feature Alignment for Remote Sensing Image Captioning." *IEEE Transactions on Geoscience & Remote Sensing* 62:1–11. <https://doi.org/10.1109/TGRS.2024.3401576>.
- Zhai, X., B. Mustafa, A. Kolesnikov, and L. Beyer. 2023. "Sigmoid Loss for Language Image Pre-training." 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France: 11941–11952. IEEE. Oct. <https://doi.org/10.1109/ICCV51070.2023.01100>.
- Zhang, Z., W. Zhang, M. Yan, X. Gao, K. Fu, and X. Sun. 2022. "Global Visual Feature and Linguistic State Guided Attention for Remote Sensing Image Captioning." *IEEE Transactions on Geoscience & Remote Sensing* 60:1–16. <https://doi.org/10.1109/TGRS.2021.3132095>.
- Zheng, X. 2025. Retrieval Augmented Generation and Understanding in Vision: A Survey and New Outlook." arXiv: arXiv:2503.18016. Mar 23. <https://doi.org/10.48550/arXiv.2503.18016>.
- Zheng, Y. 2024. Llamafactory: Unified Efficient Fine-Tuning of 100+ Language Models." arXiv: arXiv:2403.13372. <https://doi.org/10.48550/arXiv.2403.13372>.
- Zhou, Y., and G. Long. 2023. Style-Aware Contrastive Learning for Multi-Style Image Captioning In Findings of the Association for Computational Linguistics: EACL 2023, pages 2257–2267, Dubrovnik, Croatia. Association for Computational Linguistics.. arXiv: arXiv:2301.11367. Jan 26. <https://doi.org/10.48550/arXiv.2301.11367>.