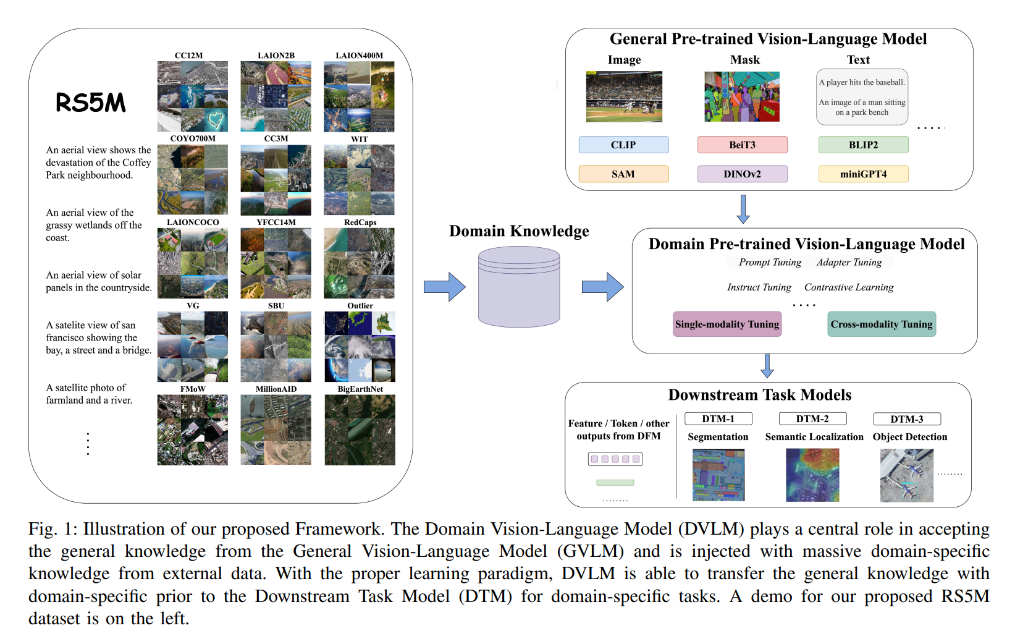
1. **Background and Problem Statement**

The emergence of Vision-Language Foundation Models provides us with more intelligent and powerful ways to tackle the usual vision and multimodality tasks. They account for the semantic meaning of the objects and their relationships, enabling reasoning about images and their associated textual descriptions, allowing for a deeper understanding of the underlying semantics [1], boosting the performance of downstream automatic analysis tasks, like semantic segmentation, object detection, image captioning, visual question answering, visual grounding, image-text retrieval, etc. They also benefit many professional fields, enabling us to solve problems from various real-world application scenarios.

When applying VLMs trained with common objects to remote sensing area, they usually underperform due to the mismatch between domains. Several works have made efforts to make use of these generalized VLMs to accomplish domain-specific downstream tasks in remote sensing context.

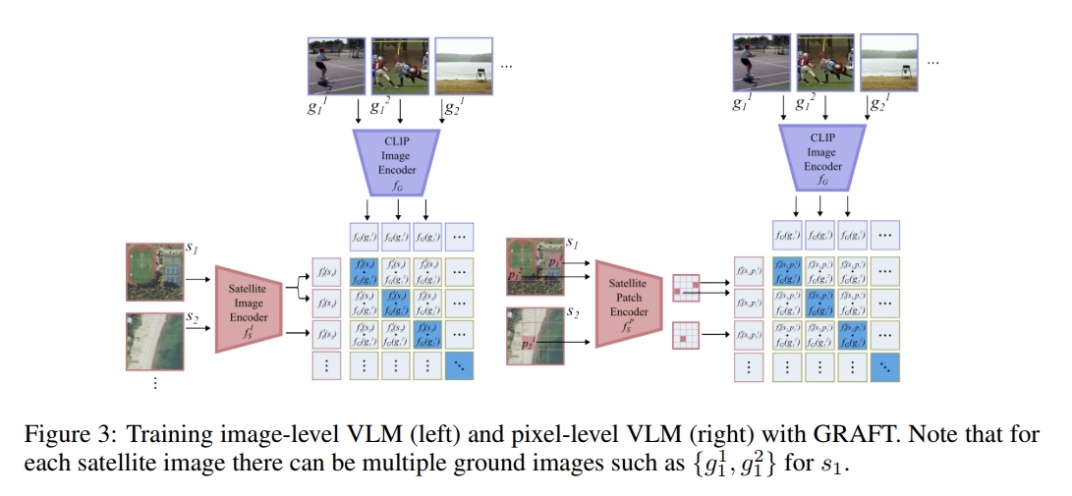
1. **Literature Reading**

* In [2], it introduces a novel framework of GVLM-DVLM-DTM to tackle the challenge of transferring both generalizability of the GVLM and domain prior knowledge to a domain-specific downstream task. Besides, to fill the gap of lacking large-scale image-text paired dataset in remote sensing area, it constructs a larger-scale dataset RS5M, which is 1000 times larger than the existing largest one. To construct this dataset, it first filters 11 public image-text paired datasets by utilizing fastdup for invalid image checking and deduplication and using VLM and the RS image Detector to clean the datasets. It then captions the filtered datasets using VLMs pre-trained on images with common objects. Four parameter-efficient fine-tuning methods are adopted to fine-tune the GVLM as DVLM candidates and demonstrate superior performance on downstream tasks in zero-shot classification and semantic localization.



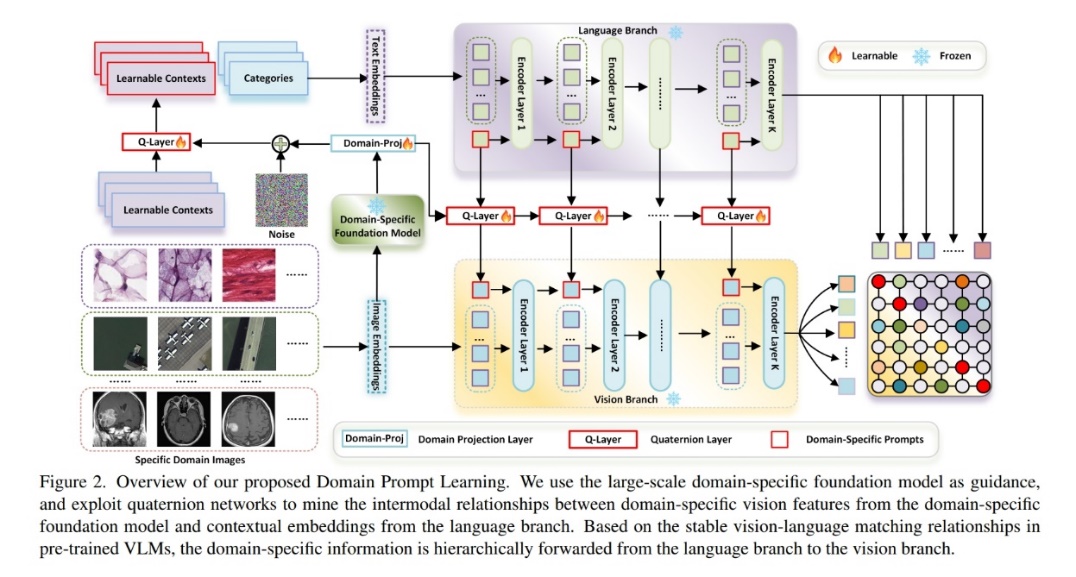
Limitation and future work: More complex DVLMs can be explored to account for the interaction between image and text , which are not fully exploited in the parameter efficient fine-tuning methods adopted in this work. Besides, other GVLM models apart from CLIP and more various downstream remote sensing tasks can also be explored.

* In [3], to build a vision-language model for satellite images without textual annotations, it utilizes internet images captured on the ground as an intermediary. To be more specific, it acquires satellite and ground image pair sets. Then to project satellite image embedding to the same representation space with CLIP at image level, it constructs a loss function based on the principle that a satellite image’s embedding should be close to all the ground images within this satellite image’s geographical scope, while stay far away from the ground images beyond this scope. It also constructs a loss function for pixel-level VLMs based on the principle that the pixel representation should stay close with its corresponding ground image. It further combines other foundational models to extend downstream tasks.



Limitation and future work: Enable the model to generate text, accomplish tasks other than VQA can be explored. Besides, since text-encoder is not fine-tuned, there may be bias in concepts of ground images, which may be mitigated by fine-tuning with a small amount of satellite image-caption data.

* In [4], to fill the gap of using domain prompt learning to transfer the recognition ability of VLMs from generalized to specialized domain, it uses domain-specific vision features from domain-specific foundation models to guide the transformation of generalized contextual embeddings from the language branch into a specialized space within the quaternion networks [4].



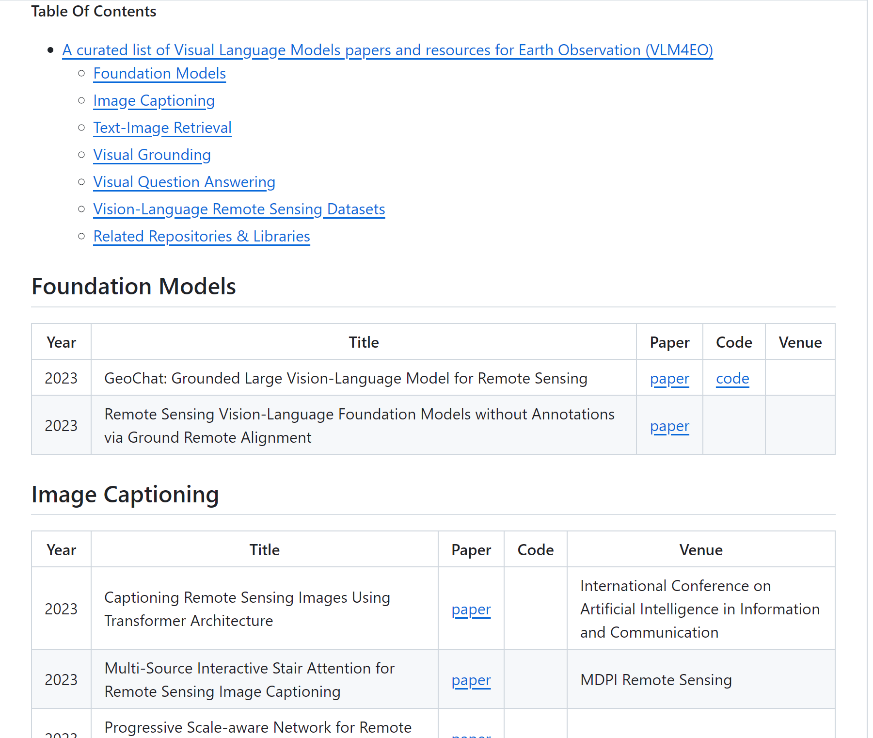
Limitation and future work: Although not specifically indicated in the literature, other network architecture or method that can promote the interaction between vision and language branch to accomplish domain prompt learning may be explored.

1. **Possible Extended Reading**

<https://github.com/Jack-bo1220/Awesome-Remote-Sensing-Foundation-Models>

<https://github.com/geoaigroup/awesome-vision-language-models-for-earth-observation>

<https://github.com/geoaigroup/awesome-vision-language-models-for-earth-observation>



**References:**

[1] Mall, Utkarsh, et al. "Remote Sensing Vision-Language Foundation Models without Annotations via Ground Remote Alignment." *arXiv preprint arXiv:2312.06960* (2023).

[2] Zhang, Zilun, et al. "RS5M: A Large Scale Vision-Language Dataset for Remote Sensing Vision-Language Foundation Model." *arXiv preprint arXiv:2306.11300* (2023).

[3] Mall, Utkarsh, et al. "Remote Sensing Vision-Language Foundation Models without Annotations via Ground Remote Alignment." *arXiv preprint arXiv:2312.06960* (2023).

[4] Cao, Qinglong, et al. "Domain Prompt Learning with Quaternion Networks." *arXiv preprint arXiv:2312.08878* (2023).