

**DOMAIN GENERALIZED REMOTE SENSING SCENE  
CAPTIONING VIA COUNTRY-LEVEL GEOGRAPHIC  
INFORMATION**

by  
KEREM AYDIN

Submitted to the Graduate School of Engineering and Natural Sciences  
in partial fulfilment of  
the requirements for the degree of Master of Science

Sabanci University  
June 2025

KEREM AYDIN 2025 ©

All Rights Reserved

## ABSTRACT

### DOMAIN GENERALIZED REMOTE SENSING SCENE CAPTIONING VIA COUNTRY-LEVEL GEOGRAPHIC INFORMATION

KEREM AYDIN

DATA SCIENCE M.Sc. THESIS, JUNE 2025

Thesis Supervisor: Prof. Erchan Aptoula

Keywords: Scene Captioning, Domain Generalization, Remote Sensing and Open  
Vocabulary Classification

This thesis investigates the impact of incorporating country-level, text-based geographical information into a large-scale vision-language model fine-tuned for captioning optical remote sensing imagery. We hypothesize that enriching visual inputs with corresponding geographical context can enhance model performance, particularly in generalizing to images from previously unseen countries or continents. To test this, we fine-tune the Large Language and Vision Assistant (LLaVA) on optical satellite images from European countries, augmenting them with textual geographical descriptions, and evaluate its performance on images from other global regions. Experiments conducted across 175 countries using the newly released SkyScript dataset reveal that even lightweight geographical context—extracted from Wikipedia—can mitigate cross-country domain shifts, leading to notable improvements in captioning accuracy. These findings highlight the potential of multimodal approaches in enhancing the geographic generalization capabilities of vision-language models.

## ÖZET

### ÜLKE DÜZEYİNDE COĞRAFİ BİLGİLERLE ALAN GENELLEŞTİRİLMİŞ UZAKTAN ALGILAMA SAHNE ALTYAZILAMA

KEREM AYDIN

VERİ BİLİMİ YÜKSEK LİSANS TEZİ, HAZİRAN 2025

Tez Danışmanı: Prof. Erchan Aptoula

Anahtar Kelimeler: Sahne Altyazılama, Alan Genellemesi, Uzaktan Algılama ve  
Açık Sözlük Sınıflandırılması

Bu tez, optik uzaktan algılama görüntülerinin altyazılanması (captioning) amacıyla ince ayar yapılan büyük ölçekli bir görsel-dil modeline ülke düzeyinde metin tabanlı coğrafi bilginin dahil edilmesinin etkisini araştırmaktadır. Görsel girdilerin karşılık gelen coğrafi bağlamla zenginleştirilmesinin, özellikle daha önce görülmemiş ülke veya kıtalardan gelen görüntülere genelleme yapma konusunda model performansını artırabileceği hipotezini öne sürüyoruz. Bu amacı test etmek için, Large Language and Vision Assistant (LLaVA) modeli Avrupa ülkelerine ait optik uydu görüntüleri ve bunlara ait metinsel coğrafi açıklamalarla birlikte ince ayarlandı ve farklı kıtalardan gelen görüntüler üzerinde değerlendirildi. 175 ülkeyi kapsayan ve yeni yayımlanan SkyScript veri kümesi üzerinde yürütülen deneyler, Wikipedia gibi kaynaklardan elde edilen basit coğrafi bağlamın bile ülkeler arası alan farklılıklarını azaltarak altyazılama doğruluğunda kayda değer iyileşmelere yol açtığını ortaya koymaktadır. Bu bulgular, görsel verilerle metinsel coğrafi bağlamın birleştirildiği çok modlu yaklaşımların, görsel-dil modellerinin farklı ve daha önce görülmemiş coğrafi bölgelerdeki genelleme yeteneklerini artırmada önemli bir potansiyele sahip olduğunu göstermektedir.



## ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my advisor, Prof. Erchan Aptoula, for his constant guidance, encouragement, and insightful feedback throughout this research. Also I would like to gratefully acknowledge the financial support from The Scientific and Technological Research Council of Türkiye (TÜBİTAK), under grant number 123R108.

Finally, I want to thank my family and friends for their endless support, love, and patience during this journey.

*To my parents, for their unwavering support and encouragement*

## Contents

<b>LIST OF TABLES</b> .....	<b>xi</b>
<b>LIST OF FIGURES</b> .....	<b>xii</b>
<b>LIST OF ABBREVIATIONS</b> .....	<b>xiv</b>
<b>1. Introduction</b> .....	<b>1</b>
1.1. Motivation and Research Questions .....	4
1.2. Contributions .....	5
1.3. Thesis Organization .....	7
<b>2. Related Works and Preliminary Concepts</b> .....	<b>8</b>
2.1. Related Works .....	8
2.1.1. Domain Adaptation .....	8
2.1.2. Domain Adaptation on Remote Sensing .....	10
2.1.3. Domain Adaptation on Remote Sensing Using VLMs .....	11
2.2. Preliminary Concepts .....	12
2.2.1. Remote Sensing .....	12
2.2.2. Domain Generalization .....	13
2.2.3. Multimodal Models .....	14
2.2.4. Visual Language Models .....	16
2.2.5. Core Components of a Visual Language Model .....	21
2.2.5.1. Vision Encoder .....	21
2.2.5.2. Embedding Models .....	22
2.2.6. LLaVA (Large Language and Vision Assistant) .....	23
2.2.6.1. LLaVA Training Stages .....	24
2.2.6.2. Inference Process in LLaVA .....	25
2.2.7. LLaMA 3.2 Vision-Language Model .....	26
2.2.7.1. Architecture Overview .....	27
2.2.7.2. Attention Network Details .....	27
2.2.7.3. Instruction Tuning and Alignment .....	28

2.2.8.	Fine-tuning Visual Language Models .....	28
2.2.8.1.	Training Data Structure .....	30
2.2.8.2.	Fine-Tuning and LoRA .....	31
2.2.8.3.	Applying LoRA in VLMs .....	32
<b>3.</b>	<b>Methodology .....</b>	<b>33</b>
3.1.	Dataset Preparation .....	33
3.1.1.	Image Files .....	34
3.1.2.	Metadata Files .....	34
3.1.3.	Caption Files .....	34
3.1.4.	Geographic Localization via Coordinates .....	35
3.1.5.	Geographic Metadata Enrichment via Wikipedia .....	35
3.2.	Training Pipeline .....	36
3.2.1.	Conversation-Style Dataset Construction .....	37
3.2.2.	Implementation Details .....	38
3.2.3.	Training Configuration .....	38
3.3.	Evaluation .....	38
3.3.1.	Inference .....	39
3.3.2.	Evaluating Different Settings .....	39
3.3.3.	Evaluation Protocol .....	40
3.3.4.	Output Format and Logging .....	42
<b>4.</b>	<b>Experiments .....</b>	<b>43</b>
4.1.	Dataset .....	43
4.2.	Settings .....	45
4.2.1.	Establishing a Baseline .....	45
4.2.2.	Multimodal Models with Geographical Metadata .....	46
4.2.3.	Hyperparameter Selection .....	46
4.3.	Architecture .....	47
4.4.	Evaluation Metric .....	48
4.5.	Results & Discussion .....	49
4.6.	Quantitative Results .....	50
4.6.1.	Breakdown by Continent .....	52
4.6.1.1.	Implementation Details .....	52
4.6.1.2.	Analysis .....	52
4.6.2.	Proximity to Training Region .....	53
4.6.2.1.	Implementation Details .....	53
4.6.2.2.	Analysis .....	53
4.7.	Qualitative Analysis .....	54

5. Conclusion and Future Work .....	60
BIBLIOGRAPHY.....	64
APPENDIX A .....	69

## LIST OF TABLES

Table 3.1. Training Configuration .....	38
Table 4.1. The captions used during the experiments. ....	44
Table 4.2. Best-performing hyperparameters across model settings. ....	47
Table 4.3. Accuracies for the baseline (B) captioning model and for the proposed approach (P) reinforced with Wikipedia geography articles.	49
Table 4.4. Accuracy of different models on SkyScript .....	50
Table 4.5. Mean accuracy improvement (%) by continent.....	52
Table 4.6. Mean improvement (%) by proximity to training region .....	53
Table 4.7. Example model outputs with and without geographical infor- mation .....	54

## LIST OF FIGURES

Figure 2.1. Remote Sensing images from UCM dataset (Mehmood et al., 2022).	12
Figure 2.2. Two aerial images of Istanbul from different camera angles and seasons (Getty Images, 2025).	13
Figure 2.3. Text, image and audio modalities as input (DeepLearning.AI, 2025d).	16
Figure 2.4. Flickr data set samples (Plummer et al., 2015)	17
Figure 2.5. Example of tasks achievable by a VQA model for remote sensing data (Lobry et al., 2020)	18
Figure 2.6. Stable Diffusion output (DeepLearning.AI, 2025c).	19
Figure 2.7. Examples of Sora in text-to-video generation. Text instructions are given to the OpenAI Sora model, and it generates three videos according to the instructions. (Liu et al., 2024c).	20
Figure 2.8. Embedding model conversion (Pinecone, 2025).	20
Figure 2.9. Dividing an image into a grid of smaller patches.	21
Figure 2.10. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset’s classes. (Radford et al., 2021).	22
Figure 2.11. Similar embedding vectors (values closer to 2 appear red, values closer to 0 appear white, and values closer to -2 appear blue)(Jalammar, 2016)	23
Figure 2.12. Diagram of a Visual Language Model	24
Figure 2.13. Examples of images and image descriptions from the Conceptual Captions dataset (Sharma et al., 2018)	25
Figure 2.14. Inference with LLaVA	26
Figure 2.15. LLaMA 3 models(Meta AI, 2024).	26

Figure 2.16. Self-attention in a sentence (DeepLearning.AI, 2025a). . . . .	28
Figure 2.17. Illustration of fine-tuning: A general-purpose model (left) provides broad predictions, while a fine-tuned model (right), trained on domain-specific data (e.g., dermatology), produces more precise and context-aware responses (DeepLearning.AI, 2025b). . . . .	29
Figure 2.18. Diagram of LoRA (Hu et al., 2022) . . . . .	31
Figure 3.1. Wikipedia scraping geography section. . . . .	36
Figure 4.1. Skyscript dataset (Wang et al., 2024) . . . . .	43
Figure 4.2. Train and test split . . . . .	44
Figure 4.3. Prompt used in baseline setting (LLaVA 1.5 and LLaMA 3.2). . . . .	45
Figure 4.4. Prompt incorporating geography metadata . . . . .	46
Figure 4.5. Architecture of the proposed multimodal classification pipeline . . . . .	47
Figure 4.6. Scores of the baseline setting on each country . . . . .	50
Figure 4.7. Scores of the proposed approach setting on each country . . . . .	51
Figure 4.8. Score differences between proposed approach and baseline on each country . . . . .	51
Figure 4.9. Mauritania orthographic projection (Wikipedia contributors, 2025b). . . . .	56
Figure 4.10. Somalia orthographic projection (Wikipedia contributors, 2025c). . . . .	58
Figure 5.1. Central Anatolia Region (Wikipedia contributors, 2025a). . . . .	60
Figure 5.2. GeoChat tasks (Kuckreja et al., 2024). . . . .	61



## LIST OF ABBREVIATIONS

<b>API</b>	Application Programming Interface .....	33
<b>CLIP</b>	Contrastive Language-Image Pre-training .....	2
<b>CSV</b>	Comma Separated Values .....	37
<b>DA</b>	Domain Adaptation .....	8
<b>GQA</b>	Grouped Query Attention .....	27
<b>JSON</b>	JavaScript Object Notation .....	35
<b>LLaMA</b>	Large Language Model Meta AI .....	5
<b>LLaVA</b>	Large Language-and-Vision Assistant .....	3
<b>LLM</b>	Large Language Model .....	14
<b>LoRA</b>	Low-Rank Adaptation .....	3
<b>RGB</b>	Red Green Blue .....	3
<b>RLHF</b>	Reinforcement Learning Human Feedback .....	28
<b>SDG</b>	Single Domain Generalization .....	1
<b>SFT</b>	Supervised Fine-tuning .....	28
<b>VLM</b>	Visual Language Model .....	11

## 1. Introduction

In machine learning, especially in real-world applications, a model’s ability to generalize beyond the data it was trained on is essential. However, a common challenge arises when there is a discrepancy between the distribution of training data (the source domain) and that of unseen testing or deployment environments (the target domain). This phenomenon is known as domain shift, and it can significantly degrade model performance when the underlying data characteristics change—such as variations in lighting, geography, resolution, or sensor type.

To address this, domain adaptation has emerged as a field of research dedicated to developing methods that reduce the impact of domain shift. These techniques typically aim to align the feature distributions of the source and target domains, thereby improving generalization performance on the latter. Domain adaptation can take various forms depending on the availability of target domain labels during training. These range from supervised and unsupervised approaches to the more challenging domain generalization setting, where the model must perform well on completely unseen domains (Tuia et al., 2016). This problem becomes particularly pronounced in remote sensing tasks such as land cover monitoring and scene description, where the significant visual diversity of global landscapes often leads to substantial distribution shifts.

While many modern methods address the more difficult but realistic issue of Single Domain Generalization (SDG), which aims to learn to generalize from a single source dataset, early attempts (Balaji et al., 2018), (Dou et al., 2019) concentrated on the scenario when several source domains are accessible during training.

Moreover, owing to their immense success with computer vision applications (Dai et al., 2024), large-scale vision language foundation models have also made a strong entry into the field of remote sensing image analysis (Liu et al., 2024a; Chen et al., 2024; Fang et al., 2023). In more detail, foundation models are typically trained on extensive datasets to acquire general-purpose knowledge and often exhibit superior performance across a wide range of visual tasks. Prompt learning is a key technique

in this context for adapting them to specific tasks. Remote sensing studies have already reported the effectiveness of prompt learning for applications such as instance segmentation and classification; notable examples include RSPrompter (Chen et al., 2024), IVP (Fang et al., 2023) and Prompt-CC (Liu et al., 2023).

Specifically, the advent of multi-modal models like CLIP (Radford et al., 2021) has revolutionized the integration of textual and image embeddings. By employing contrastive learning, these models have made it possible for text and image inputs to be compatible with each other, enabling them to be seamlessly integrated as unified representations. Besides, CLIP the Gap (Vidit et al., 2023) has shown that the domain generalization can be achieved with the addition of textual data into the input. The paper adapted an image classification model to a new domain by adding source domain prompts and domain prompts into the model input.

Several models have been subsequently developed for remote sensing tasks such as visual question answering; e.g. GeoChat (Kuckreja et al., 2024), RSGPT (Hu et al., 2023), RS-LLaVA (Bazi et al., 2024) and EagleVision (Jiang et al., 2025). These papers constructed architectures for visual question answering in remote sensing images. The paper (Zhang et al., 2023) proposed an idea for the addition of linguistic knowledge to the model at the object detection task in order to achieve domain generalization.

This thesis addresses the domain shift in scene captioning using a large-scale vision-language model. Unlike the state-of-the-art, which typically relies on target domain data or (pseudo-)labels, our approach focuses on enhancing generalization through country-level geographic metadata, specifically Wikipedia articles about the target country’s geography. We hypothesize that this contextual information, which is often available through image acquisition coordinates, provides valuable discriminative signals.

To integrate this, we adapt our model architecture into a dual-input network, separately processing and fusing visual data with geographic text. By incorporating this metadata, we aim to improve the model’s understanding of geographic contexts, boosting its ability to generalize across various domains and countries. This approach enables the model to leverage rich, context-specific information that complements the image content, thereby enhancing its scene description capabilities.

The rationale behind this design is to utilize geographic context to capture country-specific characteristics, which helps the model handle variability across domains or countries more effectively.

This thesis leverages the availability of country-specific geographic information, such

as text-based geography descriptions from Wikipedia, to address the challenges posed by domain shifts in scene captioning. Unlike traditional approaches that rely on target domain data or pseudolabels, we aim to enhance the model’s generalization capacity through contextual information derived from a country’s geographical description.

Our experiments focus on tackling the significant domain shifts present in the datasets. Specifically, we train the model using images from a subset of European countries, while the test set includes images from a diverse range of countries not seen during training. This setup emphasizes the geographical domain shift, requiring the model to generalize across regions with distinct visual characteristics.

In particular, we employ LLaVA (Liu et al., 2024b), a popular open-source text-vision model, and we integrate publicly available geographic metadata with visual features to address cross-country geographical domain shift in the context of optical scene captioning. We start by fine-tuning using LoRA a baseline multi-modal (i.e. text-vision) model on visual content and scene labels, establishing a reference point.

It was not possible to train such model without the text-vision datasets (Yuan et al., 2024; Nedungadi et al., 2024; Wang et al., 2024) published in recent years.

In this thesis, we introduce a novel approach that applies the principles of Single Domain Generalization (SDG) to multimodal scene captioning. Traditional multimodal models for domain adaptation typically focus on integrating RGB and other wavelength images, such as infrared. However, multimodality can also be achieved by combining visual and textual inputs. The rise of open-source text-vision models, such as LLaVA (Liu et al., 2024b), has made text-vision integration a key advancement in domain adaptation.

We enhance our model by incorporating country-specific geographic metadata, derived from Wikipedia articles, hypothesizing that this readily available contextual information—often accessible through image acquisition coordinates—provides valuable discriminative signals. To do so, we adapt our model to a dual-input architecture, processing and fusing visual and textual geographic information separately. This design leverages geographic context to capture country-specific scene characteristics, improving the model’s ability to handle domain variability. Our approach exemplifies unsupervised domain adaptation, relying solely on metadata in the form of country-specific geographic information.

We validate our approach using the newly published Skyscript (Wang et al., 2024) dataset, which includes images from 175 countries. Our results demonstrate the

effectiveness of incorporating country-level geographic descriptions into multimodal models to mitigate domain shift.

For validation, we train and validate the model exclusively on images from European countries and test it on images from countries around the world. The outcomes highlight the benefits of integrating geographic metadata to reduce domain shift effects.

This approach is unique in its use of Wikipedia-derived geographic metadata for mitigating geographical domain shift, a method not explored in prior research. By incorporating this metadata, we show that our geographically-aware model significantly outperforms traditional approaches in scene captioning tasks under domain shift conditions. The following sections will present our results, which substantiate the substantial improvements achieved by this approach.

## **1.1 Motivation and Research Questions**

### **Motivation**

There are a data tsunami in remote sensing images, meaning there are a lots of images taken by satellites, drones and planes however most of them are unlabeled. To unlock their full potential and make these images useful for various applications, there is a growing need for scene captioning. Without proper labels, these images lack contextual meaning, and the raw data does not transform into actionable information. On top of that training a model in a general setting is tricky, since the training dataset cannot contain whole parts of the world and it would not be cost effective to use the images from all around the world for a specific task such as disaster detection. Therefore only a subsection can be used to train a model for cost effective reasons. And the performance degradation of the model in other parts of the world can be reduced by utilizing domain adaptation.

## Research Questions

- **RQ1:** How can country-level geographic metadata (e.g., Wikipedia articles) be effectively incorporated into vision-language models for scene captioning?
- **RQ2:** Can such geographic metadata improve a model’s ability to generalize to unseen countries in the presence of geographical domain shift?
- **RQ3:** To what extent does a multimodal model that fuses visual features with textual geographic information outperform standard vision-language models in remote sensing scene captioning?
- **RQ4:** How does the proposed metadata-enhanced model perform when trained on one geographic region (e.g., Europe) and tested on visually distinct global regions?
- **RQ5:** What architectural modifications (e.g., dual-input fusion) are necessary for effectively integrating auxiliary text data into vision-language models like LLaVA or LLAMA 3.2?

## 1.2 Contributions

This thesis makes several key contributions to the field of domain adaptation in multimodal remote sensing:

- A specialized version of the SkyScript dataset has been curated to study the effects of geographical domain shift. The dataset is split such that models are trained on European countries and tested on global data, enabling a realistic evaluation of cross-country generalization.
- A novel method is proposed to mitigate domain shift by integrating country-level geographic metadata—specifically, Wikipedia articles—into the input of vision-language models. This allows models to leverage geographic context as a form of weak supervision without relying on labeled data from the target domain.
- The effectiveness of this approach is validated through a series of controlled experiments using four model configurations:

- **Baseline (LLaVA):** A standard vision-language model (LLaVA) trained only on image and task-specific prompts (e.g., “What is in this scene?”).
- **Baseline (LLaMA 3.2):** A standard vision-language model (LLaMA 3.2) trained only on image and task-specific prompts (e.g., “What is in this scene?”).
- **Proposed Method (LLaVA):** The same model enhanced with geographic metadata from Wikipedia, alongside the task-specific prompt.
- **Proposed Method (LLaMA 3.2):** A second model, LLaMA 3.2, is used to demonstrate the generalizability of the approach across different multimodal architectures.

### 1.3 Thesis Organization

The remainder of this thesis is structured as follows:

- **Chapter 2** introduces related works and preliminary concepts. The first part reviews existing research in domain adaptation, especially in the context of remote sensing, and narrows down to recent advancements that utilize visual-language models (VLMs). The second part defines essential concepts such as remote sensing, multimodal models, and visual-language model architectures. Detailed descriptions of LLaVA and LLaMA 3.2, including their components and training strategies, are also provided.
- **Chapter 3** describes the methodology used in this thesis. It outlines the dataset construction process, including image collection, metadata preparation, caption generation, and the integration of geographic information. The training pipeline, model architecture, implementation details, and evaluation procedures are presented in full detail.
- **Chapter 4** presents the experimental setup and results. This includes the definition of baseline and proposed methods, evaluation metrics, and experimental configurations. Quantitative results are compared across different settings, and qualitative analyses are used to interpret the impact of geographic metadata on scene classification performance.
- **Chapter 5** concludes the thesis and discusses potential future directions. The findings are summarized, and suggestions are made for extending the approach to finer-grained geographic metadata (e.g., regional instead of country-level), as well as for applying it to tasks beyond scene classification.



## 2. Related Works and Preliminary Concepts

### 2.1 Related Works

#### 2.1.1 Domain Adaptation

Domain adaptation (DA) addresses the problem where a model trained on a labeled *source* domain is adapted to perform well on a *target* domain with a different data distribution (Patel et al., 2015). In the typical unsupervised domain adaptation setting, only the source domain has labeled data, while the target domain is unlabeled (Wang and Deng, 2018). This domain shift can significantly degrade model performance, motivating various adaptation techniques to bridge the distribution gap.

Existing domain adaptation methods can be broadly categorized into the following approaches:

- **Invariant Feature Learning (Discrepancy Minimization):** These methods aim to learn domain-invariant feature representations by minimizing statistical differences between source and target feature distributions. For example, Deep Adaptation Networks (DAN) utilize a Multiple Kernel Maximum Mean Discrepancy (MK-MMD) loss to align distributions in the feature space (Long et al., 2015). Similarly, Correlation Alignment (CORAL) minimizes the covariance differences between domains (Sun and Saenko, 2016). Adaptive Batch Normalization (AdaBN) recomputes batch normalization statistics on the target domain to adapt the model without additional parameters (Li et al., 2016).

- **Adversarial Training:** Inspired by Generative Adversarial Networks (GANs), adversarial domain adaptation methods train feature extractors to produce representations indistinguishable across domains. Domain-Adversarial Neural Networks (DANN) introduce a gradient reversal layer to encourage domain-invariant features by confusing a domain classifier (Ganin and Lempitsky, 2016). Adversarial Discriminative Domain Adaptation (ADDA) employs separate encoders for source and target, aligning them using a discriminator in an adversarial manner (Tzeng et al., 2017).
- **Self-Training and Pseudo-Labeling:** These approaches iteratively generate pseudo-labels for unlabeled target data and retrain the model to improve target domain performance. Self-ensembling techniques, such as mean teacher models, leverage teacher-student frameworks to enforce consistency on target samples (French et al., 2018). Self-training cycles with confidence-based selection have been shown to boost adaptation effectiveness (Lee, 2013).
- **Generative (Pixel-Level) Adaptation:** GAN-based image translation methods adapt source images to resemble the target domain style, allowing standard supervised models to generalize better. Coupled GANs (CoGAN) jointly generate paired images from both domains without explicit correspondence (Liu and Tuzel, 2016). Cycle-consistent GANs (CycleGAN) enforce mappings between domains that preserve semantic content, as used in CyCADA (Hoffman et al., 2018).

Although these approaches have advanced domain adaptation, challenges remain in aligning complex, high-dimensional distributions effectively. Contemporary research often integrates multiple strategies, such as combining adversarial training with pixel-level translation, to leverage their complementary strengths.

### 2.1.2 Domain Adaptation on Remote Sensing

The application of deep learning techniques in remote sensing has significantly advanced tasks such as land cover classification, scene recognition, and semantic segmentation. However, a persistent challenge in this domain is the issue of domain shift, where models trained on data from a specific distribution (source domain) perform poorly when applied to data from a different distribution (target domain). This shift can result from variations in sensor types, resolutions, atmospheric conditions, or geographic regions.

To address domain shift, domain adaptation (DA) techniques have been developed to transfer knowledge from a labeled source domain to an unlabeled or differently distributed target domain. In remote sensing, DA methods can be categorized into several approaches. Invariant feature learning focuses on learning representations that remain consistent across domains. For example, techniques like Maximum Mean Discrepancy (MMD) (Tzeng et al., 2014) have been used to align feature distributions between source and target domains. Adversarial training, inspired by Generative Adversarial Networks (GANs), trains models to generate features that are indistinguishable across domains. Benjdira (Benjdira et al., 2019) applied GANs for unsupervised domain adaptation in semantic segmentation of aerial images, showing improvements across various urban scenes. Self-training and pseudo-labeling methods iteratively assign pseudo-labels to unlabeled target data, refining the model’s performance on the target domain. For example, the ST-DASegNet (Zhao et al., 2024) framework combines self-training with domain disentanglement to enhance cross-domain semantic segmentation. Graph Neural Networks (GNNs) have been explored in recent studies (Yuan et al., 2022), (Saha et al., 2022) for domain adaptation, utilizing their ability to model relationships between data points, which has shown promise in capturing structural similarities across domains. Lastly, generative approaches, such as CycleGAN (Zhu et al., 2017), translate images from the source domain to resemble those in the target domain, thus facilitating better model generalization.

Despite these advancements, challenges remain in effectively adapting models across diverse remote sensing datasets, especially when labeled data in the target domain is scarce or unavailable. The complexity of remote sensing data, characterized by high dimensionality and variability, necessitates continued research into robust and scalable domain adaptation methods.

### 2.1.3 Domain Adaptation on Remote Sensing Using VLMs

The integration of textual and visual data in remote sensing has gained significant attention, especially with the advent of Vision-Language Models (VLM) that bridge the gap between image data and natural language. These models have shown promise in addressing domain adaptation challenges by leveraging multimodal information.

Recent studies have explored the adaptation of large-scale VLMs, such as CLIP, to the remote sensing domain. For instance, (Zhang et al., 2024) introduced GeoRSCLIP, a large-scale dataset comprising 5 million remote sensing images paired with English descriptions, facilitating the fine-tuning of CLIP for remote sensing tasks. Their model, GeoRSCLIP, demonstrated improved performance in zero-shot classification and cross-modal retrieval tasks.

Similarly, (Silva et al., 2024) proposed RS-M-CLIP, a multilingual VLM fine-tuned on translated remote sensing datasets. By incorporating multilingual inputs, their model achieved state-of-the-art results in cross-modal and multilingual image-text retrieval tasks.

Addressing the scarcity of annotated data in remote sensing, (Mall et al., 2023) developed an unsupervised method that aligns remote sensing images with ground-level images using CLIP, eliminating the need for textual annotations. Their approach enabled zero-shot classification and segmentation tasks, outperforming supervised counterparts.

(Cha et al., 2024) introduced a method to curate vision-language datasets without human annotations by employing an image decoding model. Their model achieved superior performance in downstream tasks such as zero-shot classification and semantic localization.

The application of instruction tuning and prompt engineering has further enhanced the capabilities of VLMs in remote sensing. SkyEyeGPT, developed (Zhan et al., 2025), unified multiple remote sensing vision-language tasks through instruction tuning with a large language model. Their model demonstrated superiority across various datasets for tasks like image captioning and visual grounding.

(Zheng et al., 2025) introduced UniRS, a vision-language model unifying multi-temporal remote sensing tasks. By adopting a unified visual representation and prompt augmentation mechanism, UniRS achieved state-of-the-art performance in tasks such as visual question answering and change captioning.

To address cross-domain challenges, domain alignment techniques have been employed. The DALV framework (Tian et al., 2024) proposed by researchers at the ACM International Conference on Information and Knowledge Management utilized a dual-modality prototype guided pseudo-labeling mechanism, leveraging pre-trained VLMs like CLIP for cross-domain remote sensing image retrieval. Their approach outperformed existing methods in various retrieval tasks.

## 2.2 Preliminary Concepts

### 2.2.1 Remote Sensing



Figure 2.1 Remote Sensing images from UCM dataset (Mehmood et al., 2022).

Remote sensing images are photographs or data visualizations captured by sensors mounted on satellites, drones, or aircraft that record information about the Earth’s surface without requiring physical contact. These sensors operate across a range of spectral bands—visible, infrared, microwave, and beyond—enabling the observation of both surface-level and subsurface features. In Fig 2.1, some examples of remote sensing images can be seen.

In recent years, the advent of high-resolution sensors and increased satellite coverage has resulted in a data tsunami: vast quantities of remote sensing imagery being collected at unprecedented spatial, spectral, and temporal resolutions. However, the sheer volume of data is not inherently valuable—data alone is not equivalent to information. Without appropriate annotation, interpretation, or context, raw

imagery remains largely unusable for analytical or decision-making purposes. This highlights the urgent need for automated and scalable methods to extract meaningful information and insights from unstructured remote sensing data.

Remote sensing imagery plays a pivotal role in a wide array of applications, including land cover classification, deforestation tracking, crop monitoring, urban expansion analysis, disaster response, and climate change research. These images enable continuous, large-scale, and cost-effective monitoring of the planet, providing timely and actionable insights that support informed decision-making across scientific, governmental, and industrial domains. By transforming raw sensor data into interpretable and context-rich information, remote sensing facilitates a deeper understanding of Earth’s dynamic and interlinked systems.

### 2.2.2 Domain Generalization

Domain generalization is a subfield of transfer learning that focuses on improving a model’s ability to perform well on unseen but related domains without requiring access to data from the target domain during training. For instance, a semantic segmentation model trained on remote sensing images of France may exhibit performance degradation when tested on remote sensing images of China due to several domain changes. These variations are referred to as domain shifts. Domain generalization aims to build models that can handle such shifts without explicit adaptation to the new domain.

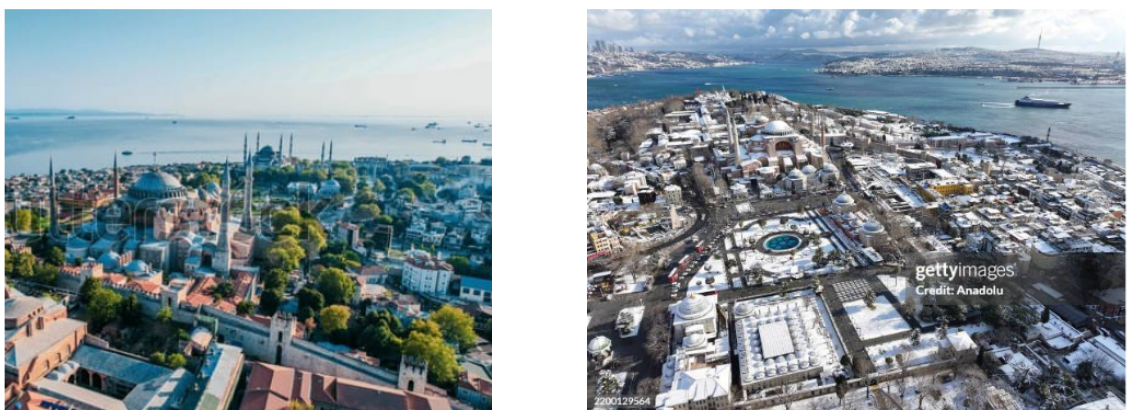


Figure 2.2 Two aerial images of Istanbul from different camera angles and seasons (Getty Images, 2025).

Figure 2.2 illustrates two aerial images of the same geographic location captured

under different conditions—one during a sunny summer day and the other in snowy winter. Although they depict the same location, the images differ significantly in visual appearance due to seasonal variation, lighting, and camera angle.

A model that relies purely on low-level visual features may struggle to recognize that both images represent the same scene, as the pixel-level distributions and texture patterns differ substantially. This exemplifies a domain shift—a core challenge in computer vision where a model trained on one distribution (e.g., summer imagery) fails to generalize to another (e.g., winter imagery).

To address this, domain generalization focuses on learning representations that are robust to such variations and can transfer across unseen domains. In our work, we highlight the importance of moving beyond surface-level visual similarities by incorporating semantic understanding. By leveraging the reasoning capabilities of large language models (LLM) through multimodal approaches, we enable models to process both visual and textual information. This fosters a deeper and more generalizable understanding of scenes, allowing the model to perform well in new domains without requiring additional task-specific training data.

### 2.2.3 Multimodal Models

In multimodal learning, the central objective is to combine information from multiple data modalities—such as text, images, audio, or video—to perform prediction or reasoning tasks that cannot be solved effectively using a single source of information. Unlike unimodal learning, which relies on one type of data, multimodal models aim to exploit the complementary strengths of different modalities to produce more robust and informative outputs.

There are three main categories of multimodal learning techniques:

- **Fusion-based approaches:** These involve encoding each modality into a shared or joint representation space, allowing the model to learn cross-modal interactions and correlations. Fusion can occur at different stages of the pipeline:
  - *Early fusion* combines raw features before modeling.
  - *Mid fusion* integrates intermediate representations.
  - *Late fusion* aggregates final predictions.

A commonly used dataset for such tasks is **Flickr30k** (Young et al., 2014), which supports image-text captioning tasks by combining visual features extracted from pre-trained CNNs with textual features such as word embeddings or bag-of-words representations.

- **Alignment-based approaches:** These seek to align the representations of different modalities to a common structure or timeline. This strategy is particularly effective when modalities are directly related and need to be mapped to one another, as in audio-visual speech recognition or sign language recognition. The **RWTH-PHOENIX-Weather 2014T** (Camgoz et al., 2018) dataset is a well-known resource in this category, containing video and audio recordings of German Sign Language (DGS) suitable for alignment-based learning.
- **Late fusion:** This technique involves combining the predictions of independently trained unimodal models, each specialized in a single modality. It is useful when the modalities provide complementary perspectives without strong direct correlation. A real-world example is emotion recognition in music using the **DEAM** dataset (Aljanaki et al., 2017), which includes both audio features and lyrics. Separate models can be trained on each modality, and their predictions can then be combined to produce a final, fused prediction.

While multimodal learning offers significant advantages, it also presents several challenges. These include:

- **Representation** – learning meaningful and compatible representations for each modality,
- **Fusion** – effectively integrating multimodal data,
- **Alignment** – mapping different modalities onto a common space or structure,
- **Translation** – generating one modality from another,
- **Co-learning** – enhancing learning in one modality through signals from another.



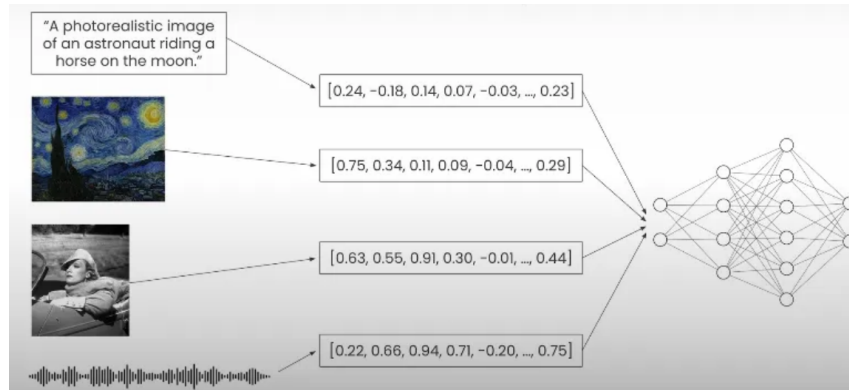


Figure 2.3 Text, image and audio modalities as input (DeepLearning.AI, 2025d).

Multimodal models are inherently flexible and can incorporate any combination of modalities, including audio, image, video, and text. Depending on the application, these models may use just two modalities (e.g., image and text in visual question answering) or multiple modalities together (e.g., video, audio, and text in multimedia content analysis). This flexibility enables a wide range of tasks and unlocks richer, more context-aware machine learning systems. In Fig 2.3, a multimodal model using 3 different modalities can be see.

#### 2.2.4 Visual Language Models

Multimodal models are a class of machine learning systems that process multiple input modalities—most commonly vision and language. These models are specifically designed to interpret, relate, and generate content across these different data types. Common applications include image captioning (e.g., CLIP), visual question answering (e.g., GPT-4-vision), and generative tasks such as text-to-image (e.g., DALL-E) or text-to-video (e.g., SORA) synthesis.

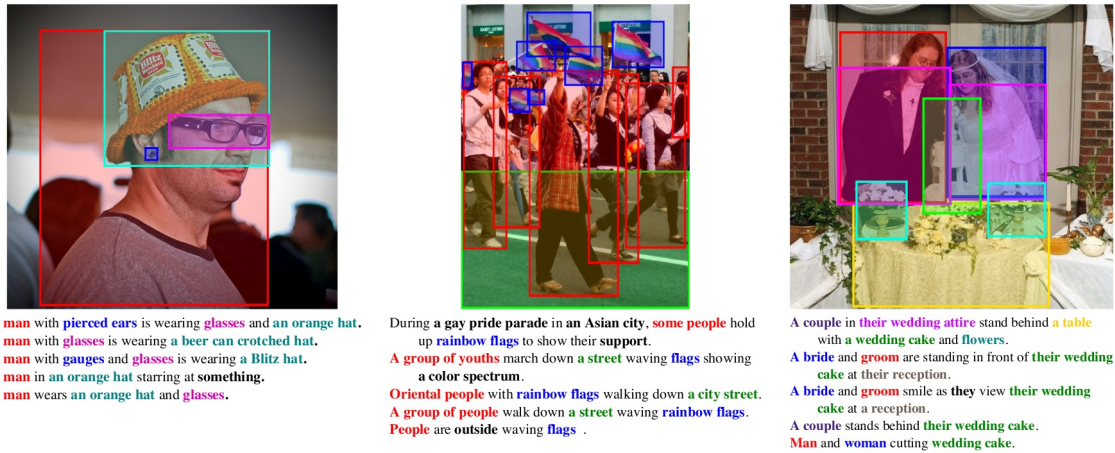


Figure 2.4 Flickr data set samples (Plummer et al., 2015)

- Image Captioning:** This task involves generating a natural language description of the content in an image. It requires the model to understand both the objects and their spatial relationships within the scene. For example, a model trained on the Flickr30k (Plummer et al., 2015) as it can be seen in Fig 2.4 or COCO dataset can take an image as input and output captions such as “A boy playing soccer in a park” or “Two dogs running through the snow.” Models like CLIP and BLIP are often used for encoding visual information in combination with language models for generation.



Classical tasks	<table><tr><th>Classification</th></tr><tr><td>Is it a rural or urban area?</td></tr></table>	Classification	Is it a rural or urban area?	<table><tr><th>Regression</th></tr><tr><td>How many buildings are there?</td></tr></table>	Regression	How many buildings are there?
	Classification					
	Is it a rural or urban area?					
Regression						
How many buildings are there?						
<table><tr><th>Detection</th></tr><tr><td>Is there a road?</td></tr></table>	Detection	Is there a road?				
Detection						
Is there a road?						
Specific tasks	<table><tr><th>Regression</th></tr><tr><td>What is the area covered by small buildings?</td></tr></table>	Regression	What is the area covered by small buildings?	<table><tr><th>Detection</th></tr><tr><td>Is there a road at the top of the image?</td></tr></table>	Detection	Is there a road at the top of the image?
	Regression					
What is the area covered by small buildings?						
Detection						
Is there a road at the top of the image?						
Mix of tasks	<table><tr><th>Regression / Detection</th></tr><tr><td>What is the number of roads next to a park?</td></tr></table>	Regression / Detection	What is the number of roads next to a park?	<table><tr><th>Detection</th></tr><tr><td>Is there a building next to a parking?</td></tr></table>	Detection	Is there a building next to a parking?
	Regression / Detection					
What is the number of roads next to a park?						
Detection						
Is there a building next to a parking?						

Figure 2.5 Example of tasks achievable by a VQA model for remote sensing data (Lobry et al., 2020)

- **Visual Question Answering (VQA):** In VQA, the model is provided with an image and a textual question related to the image, and it must generate a correct and contextually grounded answer. This task requires not only visual understanding but also reasoning across modalities. For example, given an image of a dining table and the question “How many plates are on the table?”, the model must count the relevant objects. Recent models such as GPT-4 with vision capabilities or BLIP-2 demonstrate strong performance in VQA tasks.

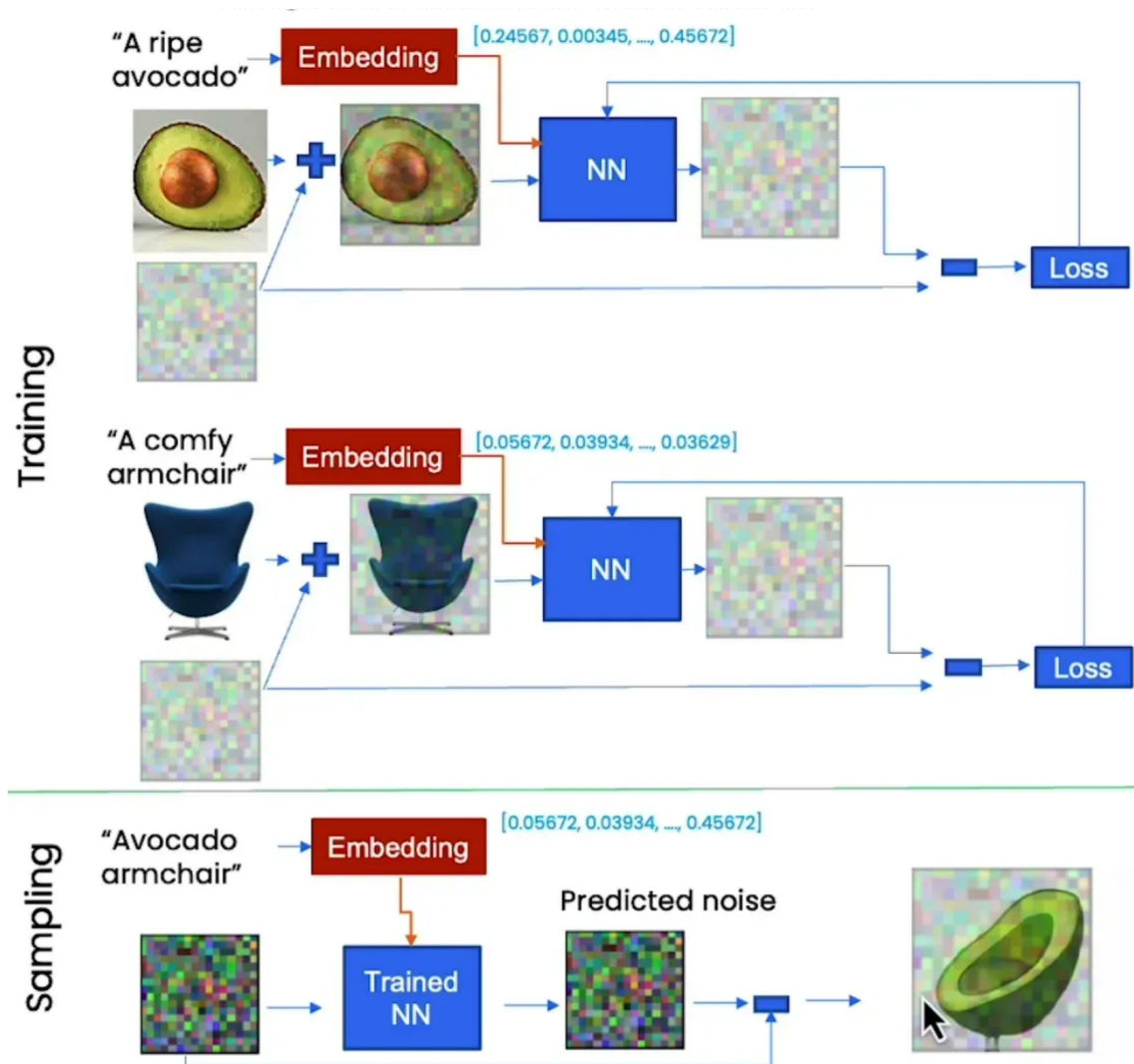


Figure 2.6 Stable Diffusion output (DeepLearning.AI, 2025c).

- **Text-to-Image Generation:** In this generative task, the model takes a textual prompt as input and generates a corresponding image. This requires learning a deep semantic correspondence between language and visual concepts. For instance, a prompt like "A futuristic city at sunset with flying cars" would be converted into a plausible image matching the description. DALL-E and Stable Diffusion are prominent models in this domain.

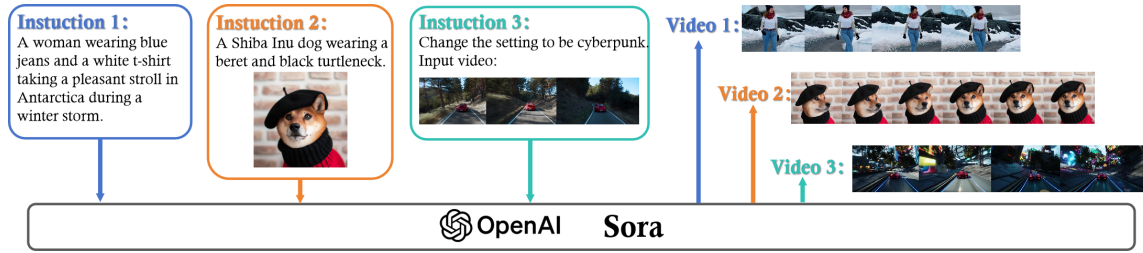


Figure 2.7 Examples of Sora in text-to-video generation. Text instructions are given to the OpenAI Sora model, and it generates three videos according to the instructions. (Liu et al., 2024c).

- **Text-to-Video Generation:** Extending the concept of text-to-image, text-to-video generation involves producing an entire video sequence from a textual description. This task is more complex as it introduces the temporal dimension—requiring models to understand motion, consistency, and coherence across frames. A prompt such as “A surfer riding a giant wave at sunset” must be translated into a dynamic, temporally smooth video. Models like **SORA** are state-of-the-art examples of text-to-video generation systems.

The architecture of visual language models (VLMs) varies depending on how and when they integrate visual and textual information. Some models perform early fusion, integrating modalities at the input level; others opt for mid-level or late fusion, combining modality-specific representations at intermediate or final layers. Conceptually, VLMs operate like large language models (LLMs) that are extended to also “understand” visual input. Regardless of the fusion strategy, these models typically transform both image and text inputs into embeddings—dense vector representations capturing semantic meaning.

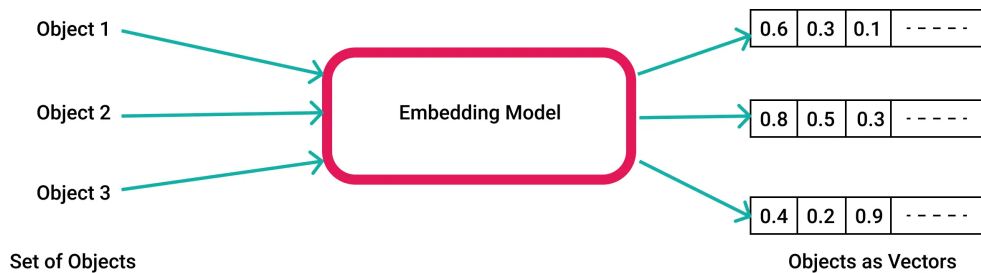


Figure 2.8 Embedding model conversion (Pinecone, 2025).

The underlying principle of VLMs is to encode both modalities into a shared semantic space, enabling unified reasoning over text and images. The next sections focus on two influential architectures in this field: LLaVA and LLaMA 3.2, both of which have significantly contributed to the progress of open-source multimodal AI.

## 2.2.5 Core Components of a Visual Language Model

Visual-language model like LLaVA and LLaMA integrate multiple components to achieve cross-modal understanding. The two foundational modules are the vision encoder and the embedding model.

### 2.2.5.1 Vision Encoder

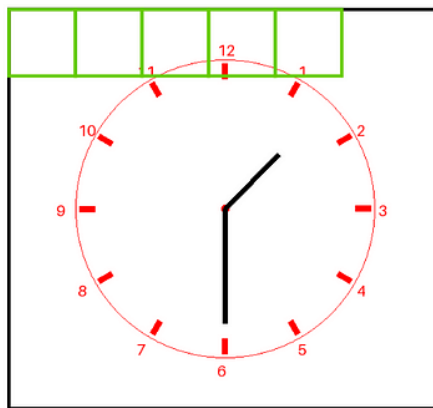


Figure 2.9 Dividing an image into a grid of smaller patches

Vision encoders often divide the image into a grid of smaller patches (typically  $16 \times 16$  pixels) and extract features from each patch in the Fig 2.9. These patches are processed into vectors that encode local visual information. The vision encoder consists of multiple layers—such as attention and feedforward layers—that gradually refine these representations through a deep network.

The ultimate goal of training the vision encoder is to produce embeddings that meaningfully correspond to language descriptions. During pre-training, the model learns to align the image and text embeddings by minimizing the distance between them in the embedding space.

For instance, the embedding of a dog image should lie close to the embedding of the word “dog” enabling the model to bridge visual and linguistic modalities in a meaningful way.

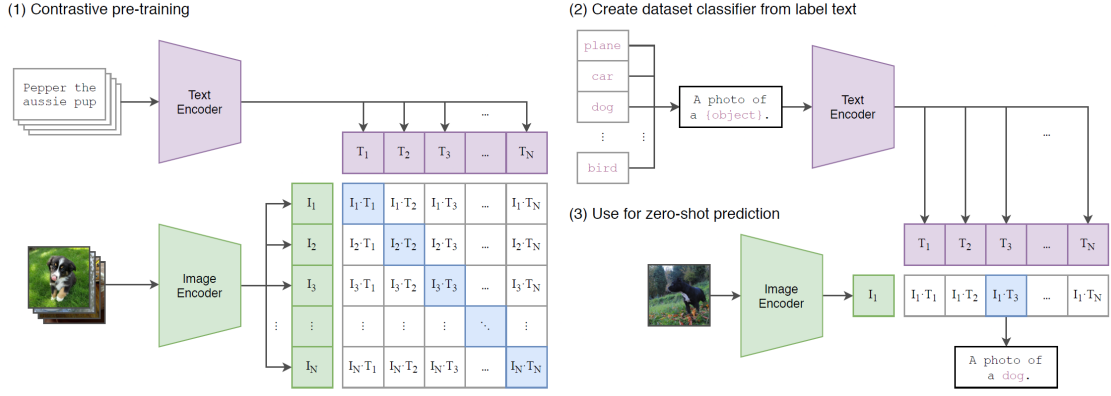


Figure 2.10 While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset’s classes. (Radford et al., 2021).

### 2.2.5.2 Embedding Models

Embedding models transform discrete inputs—such as words, tokens, or image patches—into dense, continuous vectors within a fixed-dimensional space. These embeddings enable the model to numerically represent and manipulate complex input types, making them suitable for processing by neural networks.

This approach is grounded in the *distributional hypothesis*, which posits that the meaning of a word is largely determined by its contextual usage, meaning that words appearing in similar contexts tend to share similar meanings. Embedding models capture this principle by learning from co-occurrence patterns in large corpora during training. Consequently, semantically similar inputs (e.g., "car" and "truck") are represented as nearby points within the embedding space. For instance, when visualizing the embedding vectors, cells can be color-coded based on their proximity to certain values: red for values closer to 2, white for values near 0, and blue for values closer to -2.

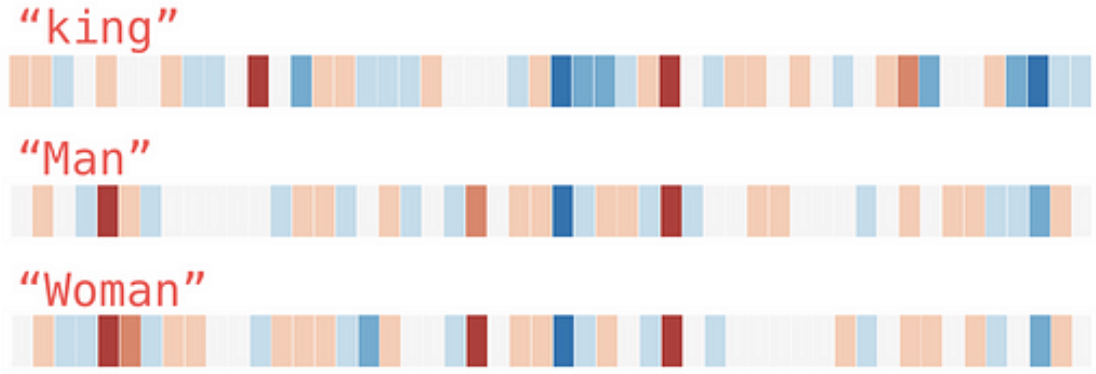


Figure 2.11 Similar embedding vectors (values closer to 2 appear red, values closer to 0 appear white, and values closer to -2 appear blue)(Jalammar, 2016)

As depicted in Fig 2.11, “Man” and “Woman” are closer to each other in the embedding space than either is to “King.” This observation indicates that the vector representations capture significant semantic relationships, highlighting the contextual and associative meaning of the words.

In the context of visual-language models (VLMs), embedding layers play a foundational role in bridging modalities. Separate embedding modules are typically used for visual inputs (e.g., using a vision encoder to embed image patches) and textual inputs (e.g., using token embeddings from a language model). These embeddings are then projected or aligned into a shared multimodal space, enabling the model to reason jointly over both types of input.

For instance, in models like CLIP or LLaVA, both image features and text tokens are mapped into the same semantic space, allowing the model to determine whether an image and a caption refer to the same concept. This shared embedding space is essential for tasks such as image captioning, visual question answering, and multi-modal classification.

Ultimately, embedding models form the basis of semantic understanding in VLMs by encoding inputs in a way that reflects both meaning and context, allowing the model to generalize effectively to new or unseen data.

### 2.2.6 LLaVA (Large Language and Vision Assistant)

LLaVA(Liu et al., 2024b) is a multimodal architecture that jointly processes visual and textual data to support diverse vision-language tasks. The model was



introduced in the paper “Visual Instruction Tuning” by researchers from Columbia University and builds on the strengths of both visual and language understanding.

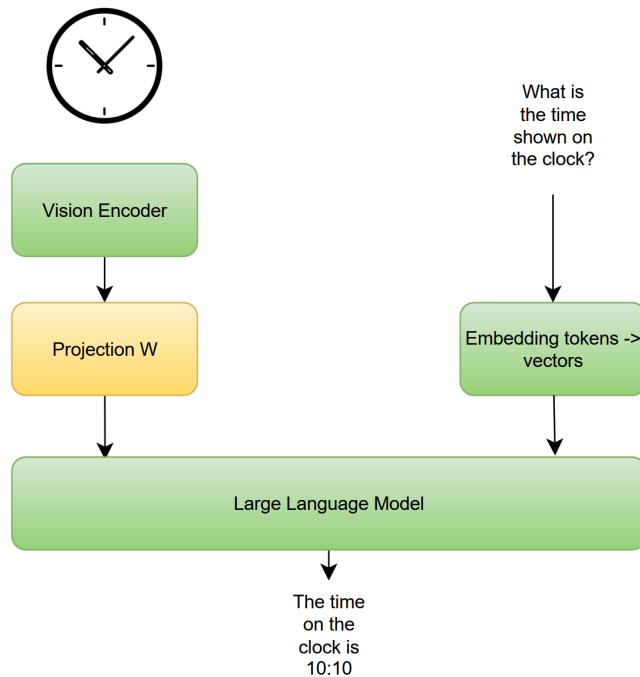


Figure 2.12 Diagram of a Visual Language Model

Its architecture combines a vision encoder with a language model, specifically Vicuna—a fine-tuned version of LLaMA 2. The visual backbone is a CLIP ViT-L/14 encoder, which is proficient in extracting rich representations from images. Vicuna, on the other hand, manages the textual inputs and interactions.

During inference, the image is first encoded into a visual embedding, while the accompanying text is processed into a textual embedding. These are then aligned within the same vector space, enabling the language model to jointly reason over both modalities.

#### 2.2.6.1 LLaVA Training Stages

LLaVA’s training involves two main stages:

- 1.1 **Feature Alignment Pre-training:** This stage focuses on learning a projection that aligns image features with their textual counterparts. It uses the CC3M dataset to fine-tune the projection layer only.



**Alt-text:** A Pakistani worker helps to clear the debris from the Taj Mahal Hotel November 7, 2005 in Balakot, Pakistan.

**Conceptual Captions:** a worker helps to clear the debris.

**Alt-text:** Musician Justin Timberlake performs at the 2017 Pilgrimage Music & Cultural Festival on September 23, 2017 in Franklin, Tennessee.

**Conceptual Captions:** pop artist performs at the festival in a city.

Figure 2.13 Examples of images and image descriptions from the Conceptual Captions dataset (Sharma et al., 2018)

**1.2 End-to-End Fine-tuning:** Both the language model and projection matrix are updated. This stage adapts the model for specific applications such as:

- **Visual Chat:** Instruction-tuned for general-purpose multimodal interactions.
- **Science QA:** Fine-tuned on domain-specific data for science-related question answering.

#### 2.2.6.2 Inference Process in LLaVA

At inference time, both image and text inputs are encoded into embeddings and fed into a large language model. The model then performs generation or classification tasks, guided by these multimodal embeddings.

LLaVA employs attention mechanisms to dynamically determine which parts of the input—visual or textual—should be prioritized during each step of the computation. This allows the model to flexibly shift its focus depending on the context, resulting in accurate and coherent responses grounded in both vision and language.

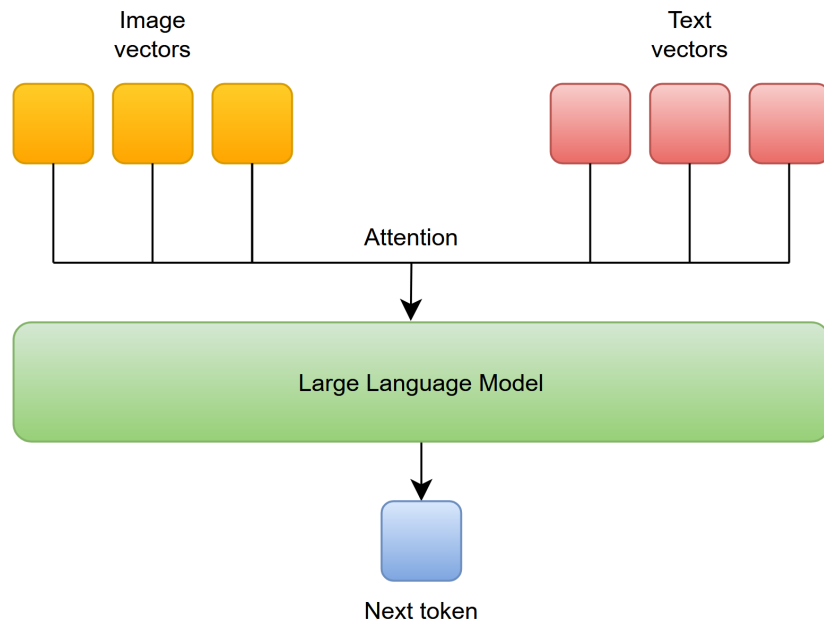


Figure 2.14 Inference with LLaVA

### 2.2.7 LLaMA 3.2 Vision-Language Model

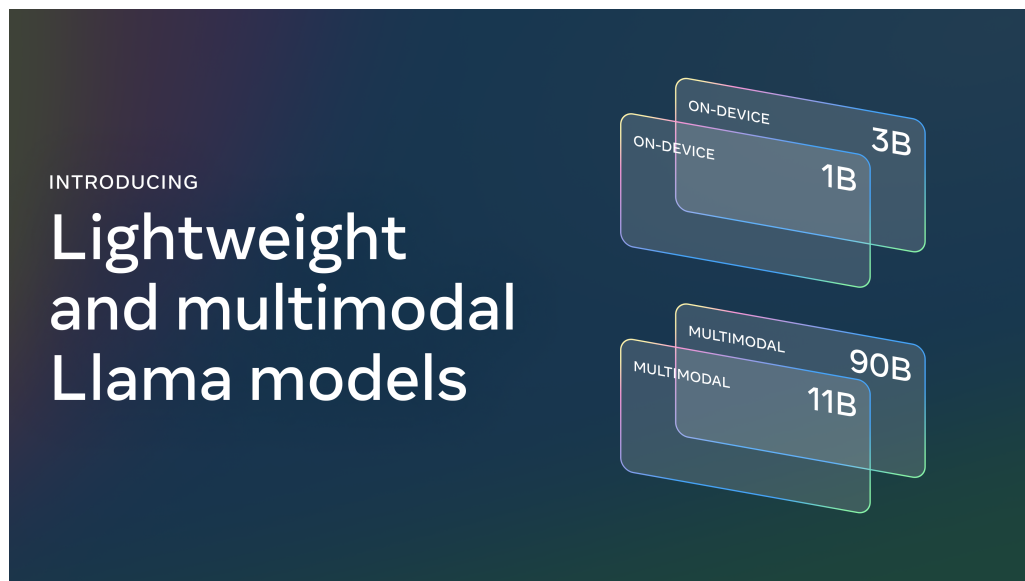


Figure 2.15 LLaMA 3 models(Meta AI, 2024).

Meta’s LLaMA 3.2 is a cutting-edge multimodal model that extends the LLaMA architecture to process both text and images, enabling a wide array of vision-language tasks. These tasks include image captioning, visual question answering, and other vision-guided reasoning tasks. LLaMA 3.2 is released in multiple variants, with the two largest models—11 billion and 90 billion parameters—incorporating vision capabilities alongside the text-only versions (1 billion and 3 billion parameters) to allow for more flexible deployment across different environments and use cases.

### 2.2.7.1 Architecture Overview

The vision encoder of LLaMA 3.2 utilizes a two-stage transformer-based approach. The first stage is a 32-layer local encoder that processes the image by dividing it into  $16 \times 16$  pixel patches, capturing important local features. The second stage consists of an 8-layer global encoder, which refines these feature representations and aggregates contextual information across the entire image using gated cross-attention mechanisms. This dual-stage design enables the model to capture both local and global visual features, essential for multimodal tasks.

**Language Model Backbone:** The textual backbone of LLaMA 3.2 is based on the LLaMA 3.1 architecture, which employs a 40-layer, decoder-only transformer. The hidden size of this backbone is 4,096, with 32 layers dedicated to self-attention and 8 layers to cross-attention. These cross-attention layers are strategically placed every five layers (at layers 3, 8, 13, 18, 23, 28, 33, and 38), creating "fusion checkpoints" where visual and textual information are integrated. This design enables the model to seamlessly combine information from both modalities during inference.

**Integration Mechanism:** Visual features from the global encoder are projected and concatenated with textual embeddings through adapter-style cross-attention layers. These layers compute attention scores between the textual input queries and the visual input keys/values, allowing the model to reference relevant image regions while generating text outputs. This interaction between vision and language embeddings is crucial for the model's ability to understand and generate content involving both modalities.

**Grouped-Query Attention (GQA):** Grouped-Query Attention (GQA) is incorporated into LLaMA 3.2 to optimize efficiency, particularly for the 90-billion parameter variant. By sharing key/value projections across multiple queries, this technique reduces computational load and enhances inference speed, all while maintaining model performance.

### 2.2.7.2 Attention Network Details

**Self-Attention:** The self-attention layers in LLaMA 3.2 allow the model to autoregressively generate text based on both the previously generated tokens and the visual context provided by the image. This enables the model to generate contextually relevant text by attending to both modalities simultaneously.

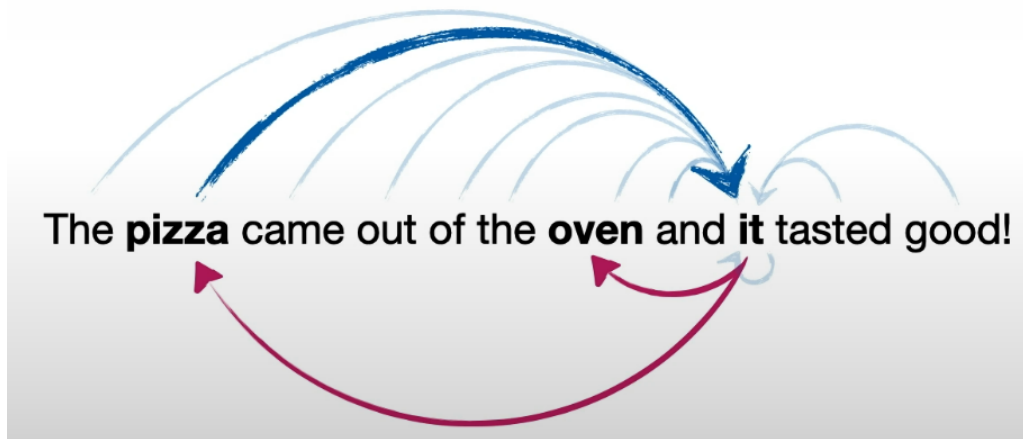


Figure 2.16 Self-attention in a sentence (DeepLearning.AI, 2025a).

**Cross-Attention:** The cross-attention layers play a critical role in multimodal tasks. At each fusion checkpoint, these layers take the textual embeddings as queries and the visual features as keys/values. This mechanism ensures that the model can focus on the most relevant parts of the image when generating corresponding text.

**Gated Fusion:** The global encoder in LLaMA 3.2 also incorporates gated attention mechanisms, which control the flow of local versus global visual features. By learning appropriate gating weights, the model can dynamically adjust which visual features are emphasized during the later stages of fusion, allowing for more refined and contextually aware representations.

### 2.2.7.3 Instruction Tuning and Alignment

LLaMA 3.2 Vision-Instruct variants are fine-tuned using both supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF) on multimodal instruction-following datasets. This process enhances the model's ability to align its outputs with user intentions, significantly improving performance on tasks like image captioning and visual question answering.

### 2.2.8 Fine-tuning Visual Language Models

Fine-tuning is a transfer learning technique where a pretrained model is adapted to a more specific task or domain using additional training data. A base model

is typically trained on a broad, diverse dataset and develops general capabilities, but may lack domain-specific nuance. Fine-tuning modifies the weights of this base model by continuing training on a smaller, task-relevant dataset, thereby aligning the model’s outputs with specialized needs.

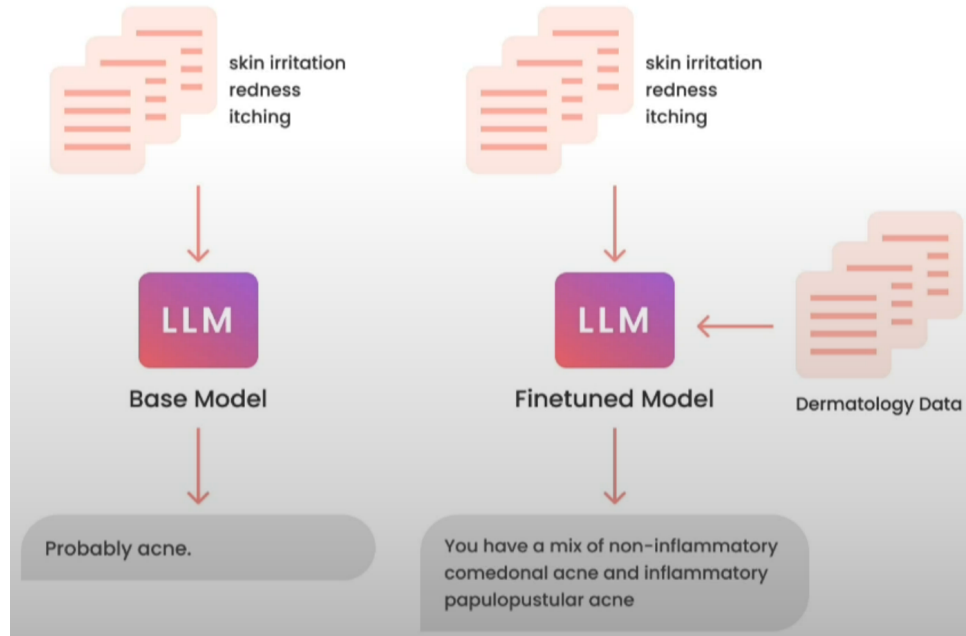


Figure 2.17 Illustration of fine-tuning: A general-purpose model (left) provides broad predictions, while a fine-tuned model (right), trained on domain-specific data (e.g., dermatology), produces more precise and context-aware responses (DeepLearning.AI, 2025b).

As illustrated in Figure 2.17, a base large language model (LLM) given inputs like “skin irritation”, “redness”, and “itching” may produce a vague diagnosis such as “Probably acne”. However, once fine-tuned on domain-specific dermatology data, the model learns to provide richer and more precise responses, such as “You have a mix of non-inflammatory comedonal acne and inflammatory papulopustular acne”.

This same principle extends to Visual Language Models (VLMs). VLMs combine a vision encoder (processing visual data) and a language model (processing text) into a unified multimodal system. Pretrained on large image-caption datasets like LAION or COCO, these models learn generalized visual-textual representations.

Fine-tuning a VLM involves further training on a smaller, task-specific dataset that includes both images and textual annotations—such as medical scans with diagnostic reports, or satellite imagery with geographic metadata. This process updates the model’s internal representations to better capture the semantics of the new domain, improving performance on specialized tasks like:

- Scene classification in specific geographic regions

- Visual question answering for medical imagery
- Multimodal document understanding

Ultimately, fine-tuning transforms a general-purpose VLM into a domain expert by incorporating relevant knowledge into its representations, improving both the accuracy and contextual richness of its outputs.

### 2.2.8.1 Training Data Structure

The training data used for VLMs generally consists of structured examples where each instance combines an image and one or more natural language prompts or interactions. One common format used for fine-tuning and instruction tuning looks like the following:

Listing 2.1 Example JSON conversation data

```
{
  "image": "images6/n1567088198_S2_18.jpg",
  "conversations": [
    {
      "from": "human",
      "value": "<image>\nClassify the image with a single
              label from [...] You must provide a single label."
    },
    {
      "from": "gpt",
      "value": "place of village"
    }
  ]
}
```

This structure emulates a dialogue between a user and an assistant (often a large language model with visual input capabilities), where the image is implicitly referenced via the `<image>` token and the task is given in natural language. The model is trained to respond with a coherent and accurate text response based on the image and prompt. This “chat-style” fine-tuning has become increasingly popular in instruction-tuned VLMs like LLaVA and MiniGPT-4, as it aligns well with the growing trend toward conversational AI.

### 2.2.8.2 Fine-Tuning and LoRA

Fine-tuning is the process of updating a pretrained model's parameters using new, typically smaller-scale, task-specific data. This allows the model to adapt its representations and decision boundaries to better fit the new domain. However, full fine-tuning of large VLMs is often computationally expensive and memory-intensive, especially when dealing with billions of parameters across vision and language backbones.

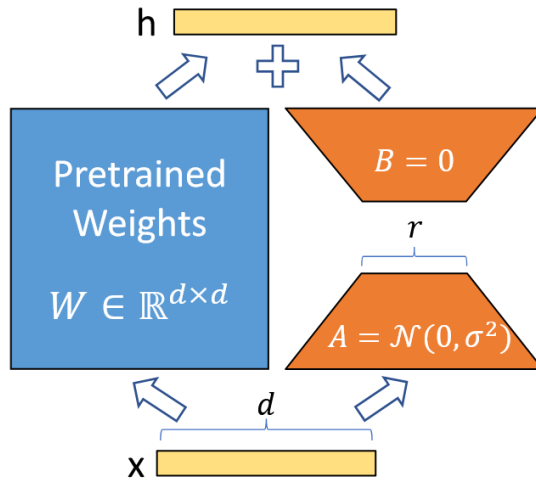


Figure 2.18 Diagram of LoRA (Hu et al., 2022)

To address this, **LoRA** (Hu et al., 2022) (Low-Rank Adaptation) offers a lightweight alternative. Instead of updating the full weight matrices during backpropagation, LoRA introduces additional trainable layers that approximate the required parameter updates using low-rank decompositions. In practice, small matrices are inserted into attention or projection layers, and only these are trained while the base model remains frozen. This drastically reduces the number of parameters that need to be updated and stored, making it possible to fine-tune large models on consumer hardware and deploy multiple task-specific adapters efficiently.



### 2.2.8.3 Applying LoRA in VLMs

When applying LoRA to a VLM, developers can selectively target specific layers: the visual encoder (e.g., ViT), the language decoder (e.g., LLaMA), or more crucially, the cross-modal fusion layers that integrate image and text information. For instance, in a classification task like the one described above, LoRA might focus on improving how the model maps visual features to relevant textual categories, without needing to relearn all of vision or language.

### 3. Methodology

In this chapter, we describe the complete methodology used to build a multimodal scene captioning pipeline using high-resolution aerial imagery and natural language descriptions. The process includes data acquisition, preprocessing, metadata integration, and prompt construction for multimodal transformer-based models. There are 3 stages:

1. Dataset Preparation
2. Training Pipeline
3. Evaluation

#### 3.1 Dataset Preparation

The initial step involved downloading the image-text pairs from the SkyScript dataset (Wang et al., 2024). Each sample includes a textual caption along with geographic coordinates (latitude and longitude). Using these coordinates, we identified the country associated with each image. To incorporate geographic metadata, we employed web scraping via the Wikipedia API to extract the “Geography” section from the corresponding country’s Wikipedia page.

We filtered the dataset to remove infrequent captions and reduce the label space from 29,000 to 37 frequently occurring captions, ensuring sufficient sample sizes per class.

### 3.1.1 Image Files

Image files are divided into six parts, named `images2.zip` through `images7.zip`, and downloaded using standard file transfer commands. These files contain the raw aerial imagery in JPEG format. They are later matched with metadata and captions for training.

### 3.1.2 Metadata Files

Each image has a corresponding metadata file in Pickle format, including:

- `box` — bounding box in (W, S, E, N) format
- `time` — timestamp as a tuple
- `center_tags` and `surrounding_tags` — semantic annotations

These metadata enrich the image data with spatial, semantic, and temporal context.

### 3.1.3 Caption Files

Filtered caption CSV files contain:

- `filepath`
- `title`
- `title_multi_objects`

These captions were polished using CLIP-based similarity filtering and ChatGPT language refinement, ensuring fluency and relevance.

### 3.1.4 Geographic Localization via Coordinates

The bounding box of each image was used to determine the country via reverse geocoding. The central coordinate of each bounding box was passed to OpenStreetMap’s Nominatim API.

---

**Algorithm 1** Determine Country from Bounding Box

---

**Require:** Bounding box ( $x_1, y_1, x_2, y_2$ )

- 1: Compute center:  $\text{lon} = (x_1 + x_2) / 2$ ,  $\text{lat} = (y_1 + y_2) / 2$
  - 2: Query geocoding service with ( $\text{lat}, \text{lon}$ )
  - 3: Extract country name from returned address
  - 4: **return** country name
- 

This procedure allowed us to determine the country for a large number of images in the 30K filtered subset. Results were saved in a text file for use in later stages.

### 3.1.5 Geographic Metadata Enrichment via Wikipedia

To further enrich the data, we scraped the “Geography” section of each country’s Wikipedia page. These geographic summaries help the model disambiguate visually similar scenes from different countries.

---

**Algorithm 2** Extract Geography Section from Wikipedia

---

**Require:** List of country names

- 1: **for** each country **do**
  - 2:   Retrieve Wikipedia page for country
  - 3:   Extract the “Geography” section
  - 4:   Clean text (remove headers, formatting artifacts)
  - 5:   Save to dictionary `geo_info[country]`
  - 6: **end for**
  - 7: Export `geo_info` dictionary to JSON
- 

An example of the scraped output is shown in Figure 3.1.

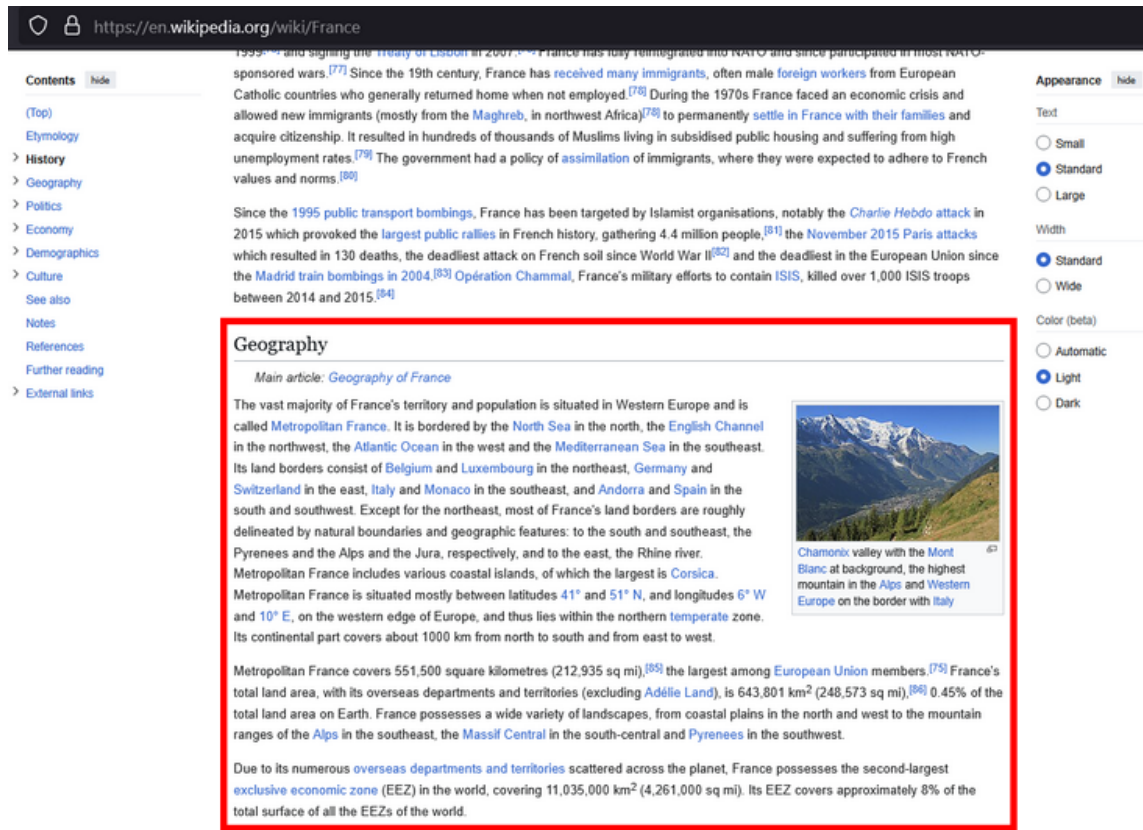


Figure 3.1 Wikipedia scraping geography section

## 3.2 Training Pipeline

This section outlines the complete training pipeline employed for fine-tuning a multimodal vision-language model to perform remote sensing scene captioning. The process involves constructing conversation-style training datasets, selecting an appropriate model architecture, implementing LoRA-based fine-tuning, and configuring training with detailed hyperparameters. Each stage is described below.

We fine-tuned a vision-language model using conversation-style datasets. Two versions of the training data were created: one without geographic metadata, and one enriched with it. These help us assess the value of location-specific context in scene captioning.

### 3.2.1 Conversation-Style Dataset Construction

To align the training process with instruction-following paradigms common in recent language models, the dataset was reformatted into a conversational style. Each image-caption pair was transformed into a multi-turn dialogue, consisting of a human-issued instruction and an expected model response.

The captioning task was evaluated under two distinct settings to compare the impact of geographic metadata:

**Setting 1: Visual-Only Captioning** — The prompt only refers to the image content.

---

**Algorithm 3** Generate Visual-Only Training Dataset

---

**Require:** Caption CSV files (train and validation)

- 1: **for** each image **do**
  - 2:   Create prompt: “Classify the image with a single label from this set: labels.”
  - 3:   Format as JSON with image and ground truth caption
  - 4: **end for**
  - 5: Shuffle dataset and save as `df_train.json` and `df_val.json`
- 

**Setting 2: Captioning with Geographic Context** — Adds Wikipedia summaries to the prompt.

---

**Algorithm 4** Generate Geo-Aware Training Dataset

---

**Require:** Caption CSVs and geographic summary JSON

- 1: **for** each image **do**
  - 2:   Retrieve country and corresponding geographic summary
  - 3:   Create prompt including geographic information
  - 4:   Format as JSON with image and target label
  - 5: **end for**
  - 6: Shuffle and save as `df_geo_train.json` and `df_geo_val.json`
-

### 3.2.2 Implementation Details

The model was trained using Hugging Face’s `transformers` library and a custom trainer `LLaVATrainer`. It supports:

- Lazy loading and shuffling
- LoRA adapters for low-rank training
- Mixed-precision computation
- CLIP-based image preprocessing

### 3.2.3 Training Configuration

Table 3.1 Training Configuration

Parameter	Value
Model	liuhaotian/llava-v1.5-13b
Vision Tower	openai/clip-vit-large-patch14
LoRA Enabled	True
Precision	bf16
Train Epochs	3
Train Batch Size	8
Learning Rate	2e-3
Scheduler	Cosine Annealing
Eval Strategy	Every 100 steps

## 3.3 Evaluation

This section outlines the procedures used to evaluate the performance of our fine-tuned multimodal models. We begin by detailing the inference process, which describes how predictions are generated from image inputs with or without auxiliary geographic metadata. Following this, we compare the model’s performance un-

der two distinct prompting strategies—visual-only and geo-aware—by conducting controlled experiments that isolate the effect of geographic context on captioning accuracy. To ensure a fair and reproducible assessment, we describe a structured evaluation protocol involving standardized sampling, prompt formatting, and decoding strategies. Given the generative nature of our model’s outputs, we employ ROUGE-1 recall as the primary evaluation metric, focusing on semantic coverage rather than exact string matches. We also introduce a strict threshold to define correctness, ensuring that only high-fidelity captions are counted as accurate. Lastly, we describe how evaluation outputs are stored and organized for further analysis. In the subsections below, these stages are explained in detail, accompanied by illustrative code snippets.

### 3.3.1 Inference

After training, the fine-tuned model is capable of generating predictions based on image inputs and optional geographic metadata. The inference process includes:

- Embedding the image via the vision encoder
- Formatting a prompt with (or without) geographic metadata
- Tokenizing and generating a response using the model

---

#### **Algorithm 5** Model Inference

---

**Require:** Image path, formatted prompt

- 1: Encode image using vision tower
  - 2: Tokenize the prompt
  - 3: Generate response using the model
  - 4: Decode and return final output
- 

### 3.3.2 Evaluating Different Settings

By running separate experiments on both training settings, we compare the performance of the model in terms of its captioning accuracy and sensitivity to geographic priors.



To assess the effectiveness of geographic metadata on image-based scene captioning, we evaluated the fine-tuned models on a per-country basis across two settings:

1. **Visual-Only Setting:** The model was prompted solely with the image content.
2. **Geographically-Informed Setting:** The model was prompted with both the image and country-specific geographic context obtained from Wikipedia (see Section 3.1.5).

### 3.3.3 Evaluation Protocol

To assess the performance of our scene captioning model, we conducted an evaluation at the country level. For each country in the dataset, a subset of up to 100 images was sampled from the full set provided in `full_dataframe.csv`. This cap was introduced to maintain evaluation consistency and reduce computational overhead, especially for countries with large image counts.

Each image was passed through an inference pipeline using both the visual-only and geo-aware model variants. For each image, a prompt was constructed in one of two formats:

- **Visual-only prompt:**  
`<image>`  
Classify the image with a single label from this set:  
`{label_list}`. You must provide a single label.
- **Geo-aware prompt:**  
`<image>`  
`<geographic_information>{Wikipedia summary}</geographic_information>`  
Classify the image according to both the image and its  
geographic information with a single label from this set:  
`{label_list}`. You must provide a single label. Explain your  
reasoning.

These prompts were passed to the model using fine-tuned weights specific to each setting. Inference was conducted with the following standardized procedure:

1. Load the fine-tuned model and tokenizer configuration.
2. Preprocess each image using the visual encoder.

3. Embed the image into the model and format the prompt accordingly.
4. Perform generation using greedy decoding (`temperature = 0.01`) to encourage deterministic outputs.
5. Compare the model’s output with the ground truth caption using ROUGE-1 recall.

### Evaluation Metric: ROUGE-1 Recall

Unlike traditional classification settings where models produce discrete labels, our approach uses instruction-following large language models that generate full natural language responses. As such, exact string matching is inappropriate. Instead, we use **ROUGE-1 recall** as our evaluation metric, which measures the overlap of unigrams (single words) between the generated output and the ground truth caption.

We chose recall (rather than precision or F1) to emphasize coverage of the reference caption. In other words, we prioritize the model mentioning all key terms in the expected description, even if the generation includes additional explanatory tokens.

To ensure reliable and interpretable predictions, we defined a strict correctness criterion: **A prediction is considered correct only if the ROUGE-1 recall exceeds 75%.**

### Caption Examples and ROUGE Computation

Ground truth captions in our dataset typically consist of compositional labels such as:

- `leisure land of pitch, landuse of retail, landuse of residential`
- `road of turning loop, landuse of industrial`
- `natural water, landuse of forest`

The model may respond with a sentence such as:

“The image contains a turning loop and is surrounded by industrial land use.”

While this is not an exact match to the label string `road of turning loop, landuse of industrial`, a ROUGE-1 recall above 75% would indicate that the model captured the core concepts correctly.

In contrast, a response like:

“This appears to be a residential area with some roads.”

would likely yield a lower ROUGE-1 recall if the target caption was more specific (e.g., mentioning “industrial” or “turning loop”), and would be marked as incorrect under our threshold policy.

## Rationale for Generative Evaluation

This evaluation protocol aligns with the goals of instruction-tuned vision-language models, where models are expected not only to classify but also to provide reasoning. By using a thresholded ROUGE-1 recall score instead of label classification accuracy, we account for semantically correct generations that may differ in syntax or order but still contain the relevant concepts. This also allows us to better evaluate the utility of geographic metadata in guiding more reasoned, context-aware predictions.

### 3.3.4 Output Format and Logging

Evaluation results are stored in:

- `scores.csv` — visual-only model results
- `geo_scores.csv` — geo-aware model results

Each file contains the number of correct/incorrect predictions per country.

## 4. Experiments

The main objective of the experiments is to reveal the positive effects of geographic metadata in the scene captioning on remote sensing images for different domains. Two runs were realized, one with a baseline LLaVA model with no geographic metadata and a model with geographic metadata. The metric used to evaluate the models is the accuracy score.

### 4.1 Dataset

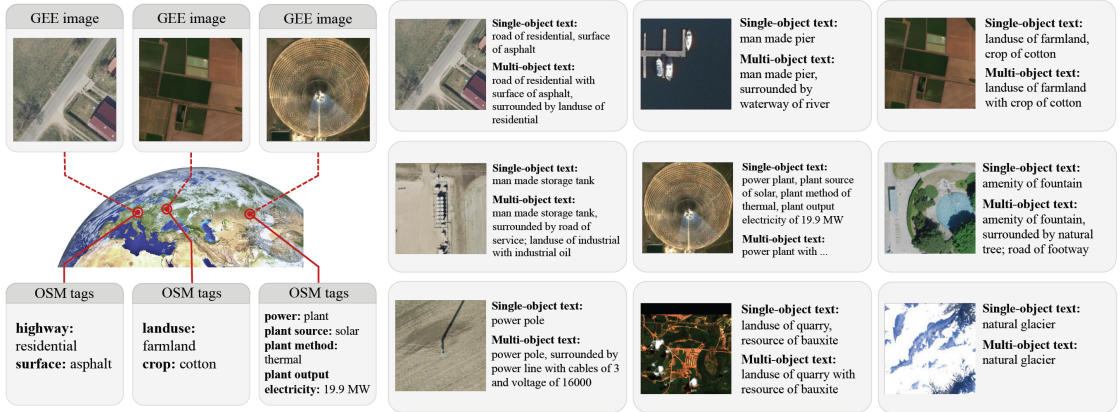


Figure 4.1 Skyscript dataset (Wang et al., 2024)

SkyScript (Wang et al., 2024) is a comprehensive vision-language dataset specifically designed for remote sensing images. As it encompasses 175 countries, it exhibits a significant amount of domain shift across countries and continents. It was constructed to address the absence of a large-scale, semantically diverse image-text dataset required for developing VLMs for remote sensing images. Unlike natural images, remote sensing images and their associated text descriptions cannot be efficiently collected from the public Internet at scale. The dataset is constructed

by using geo-coordinates to automatically connect open, unlabeled remote sensing images with the rich semantic information available in OpenStreetMap.

SkyScript comprises of 5.2 million image-text pairs and covers 29,000 distinct semantic tags from 175 countries. In the present study, only a small subset of the dataset was extracted for scene captioning, since there is a significant number of unique captions in the dataset. Only the captions that have a frequency over 15,000 were used. The number 15,000 was determined empirically, because lower numbers had an excessive number of classes. The model was trained on the task of scene captioning with 37 unique captions (Table 4.1).

Table 4.1 The captions used during the experiments.

leisure land of pitch	landuse of retail	landuse of residential
road of turning loop	natural water	landuse of commercial
airport of taxiway	highway of freeway junction	tunnel of culvert
landuse of meadow	man made pier	natural wetland
road of stop	golf hole	railway of level crossing
power tower	road of turning circle	road of service
power switch	leisure land of park	road of crossing
landuse of farmland	waterway of canal	road of residential
waterway of river	building of residential	power pole
place of village	natural peak	place of hamlet
landuse of farmyard	building of house	natural tree
building of farm	building of barn	amenity of parking space
amenity of school		

Training data consists of images from Germany, France, Switzerland, Spain, and Finland. And the testing set consists of the remaining countries all over the world. Figure 4.2 represents images of countries that are in the training dataset as red and countries that are in test dataset as yellow. As it can be seen Skyscript dataset includes images from all around the world exception of a few countries.

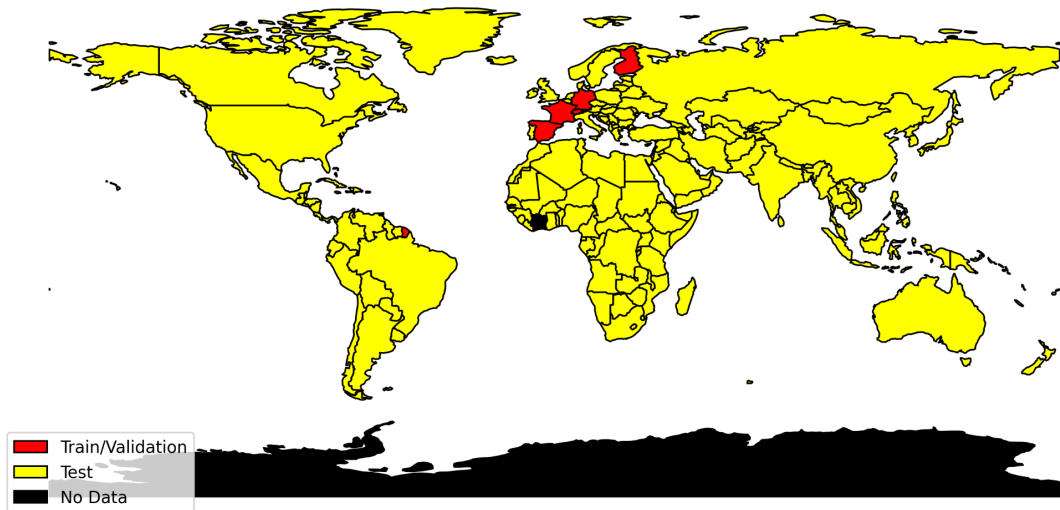


Figure 4.2 Train and test split

## 4.2 Settings


To evaluate the effectiveness of incorporating geographical metadata into multi-modal scene classification models, four distinct training settings were established:

- **Setting 1:** Baseline using LLaVA 1.5 without any metadata.
- **Setting 2:** LLaVA 1.5 with integrated geographic metadata.
- **Setting 3:** Baseline using LLaMA 3.2 without any metadata.
- **Setting 4:** LLaMA 3.2 with integrated geographic metadata.

Each model was fine-tuned for the downstream task of scene classification using adapted prompts. The core experimental configuration involved the use of cosine learning rate scheduling, Adam optimizer, and a batch size of 8 across all experiments. Hyperparameter choices are detailed in Table 4.2.

### 4.2.1 Establishing a Baseline

In the baseline settings (Setting 1 and Setting 3), the input to the model consisted only of the classification question and possible answers, without any contextual metadata. Both LLaVA 1.5 and LLaMA 3.2 were trained separately under this setting. Prompt structure followed the format proposed in (Roberts et al., 2024), modified slightly to fit the classification context of satellite imagery.



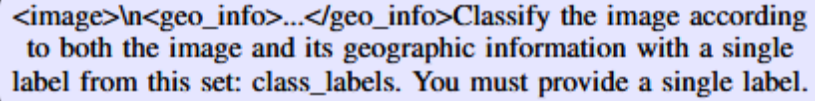
`<image> \n Classify the image with a single label from  
this set: class_labels. You must provide a single label.`

Figure 4.3 Prompt used in baseline setting (LLaVA 1.5 and LLaMA 3.2)

Training for each baseline configuration was conducted on a single NVIDIA A100 GPU over the course of one week.

### 4.2.2 Multimodal Models with Geographical Metadata

In Settings 2 and 4, the classification prompt was enriched with location-specific metadata retrieved from Wikipedia. This metadata provided contextual cues such as continent, country, and relevant textual descriptions associated with the scene’s coordinates.



```
<image>\n<geo_info>...</geo_info>Classify the image according  
to both the image and its geographic information with a single  
label from this set: class_labels. You must provide a single label.
```

Figure 4.4 Prompt incorporating geography metadata

The goal of these settings was to assess whether semantic cues from geographic knowledge could help reduce the domain shift in remote sensing tasks by providing a more informative textual context.

### 4.2.3 Hyperparameter Selection

Training was conducted on an NVIDIA A100 GPU with 80 GB of VRAM. All models were fine-tuned using the Hugging Face Transformers library, leveraging PEFT for efficient parameter-efficient fine-tuning using LoRA. In line with best practices, only the low-rank matrices and bias terms of the LoRA layers were updated, while the rest of the model parameters were frozen.

To determine optimal hyperparameters, we conducted a grid search over multiple learning rates and batch sizes. Specifically, learning rates of  $2e-3$ ,  $2e-4$ , and  $2e-5$  were evaluated alongside batch sizes of 4, 8, and 16. Each configuration was trained until convergence, with training durations ranging between 5 and 7 days per experiment. Performance was monitored on the validation set, and the configuration achieving the best evaluation score for each model variant is reported in Table 4.2.

Table 4.2 Best-performing hyperparameters across model settings.

Hyperparameter	LLaVA	LLaVA + Meta	LLaMA 3.2 Both
Learning Rate	2e-3	2e-4	2e-4
Batch Size	8	8	8
LR Scheduler	Cosine	Cosine	Cosine
Optimizer	Adam	Adam	Adam
LoRA Enabled	Yes	Yes	Yes
Training Time (approx.)	1 week	1 week	1 week

### 4.3 Architecture

To perform multimodal scene classification with and without geographic metadata, we designed a unified architecture built upon a vision-language pipeline (Figure 4.5). The model combines a vision encoder, a text encoder, and a large language model (LLM) to generate meaningful scene labels from satellite images.

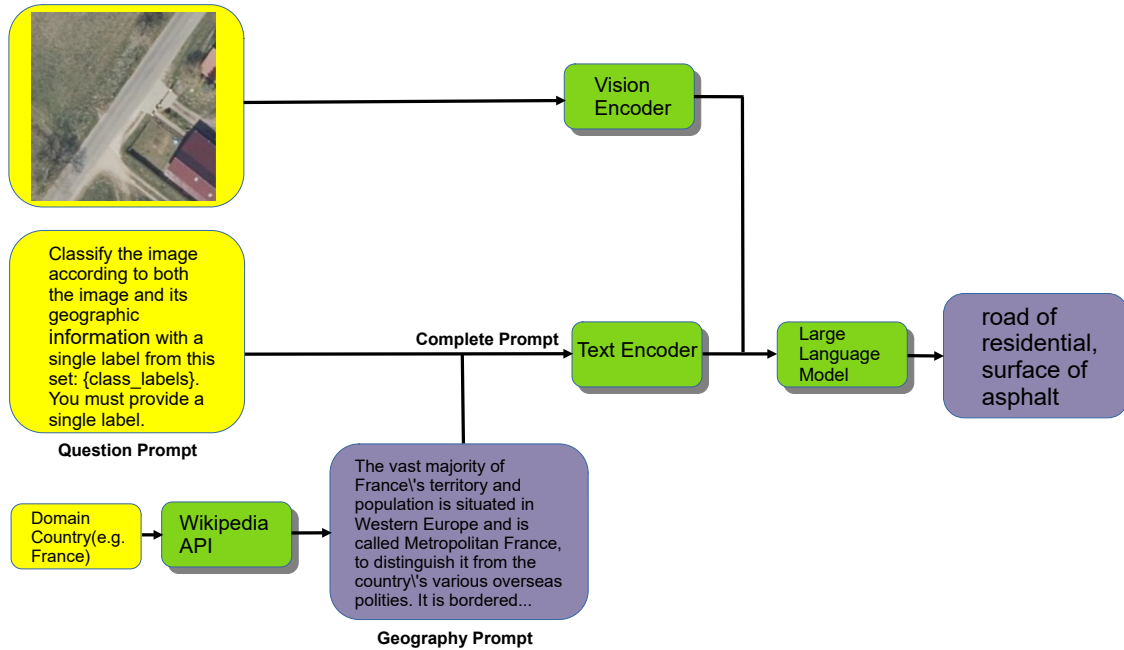


Figure 4.5 Architecture of the proposed multimodal classification pipeline

The system operates in the following stages:

1. **Image Input:** A satellite image is provided as input, which is preprocessed and passed through a pretrained vision encoder (e.g., CLIP-ViT) to extract dense image embeddings.
2. **Prompt Construction:** The input prompt includes a classification ques-



tion, a set of class labels, and—if the setting includes metadata—a geographic summary retrieved from Wikipedia. The summary is based on the country associated with the image’s coordinates.

3. **Text Encoding:** The constructed prompt is processed by a text encoder (e.g., RoPE positional encoding and tokenizer stack from LLaMA or LLaVA), transforming the tokenized input into a compatible feature space for fusion with visual tokens.
4. **Fusion and Inference:** A large language model (LLaVA 1.5 or LLaMA 3.2) receives both the visual embeddings from the vision encoder and the token embeddings from the text encoder. It processes these jointly to predict the most probable scene label from the provided set.
5. **Output:** The model returns a single label (e.g., road of residential, surface of asphalt) that best represents the input image in the given geographic context.

This architecture is capable of leveraging geographic priors to reduce ambiguity in image interpretation—especially in cases where similar visual features may correspond to different classes across regions. In settings where geographic metadata is included, the prompt is enriched with a country-specific text retrieved via the Wikipedia API. This approach supports domain adaptation without retraining on each country’s data distribution.

#### 4.4 Evaluation Metric

Since the goal is scene classification, accuracy was chosen as the evaluation metric. And since Skyscript paper used accuracy as their metric, we have chosen accuracy to do comparison at the end.

## 4.5 Results & Discussion

Table 4.3 Accuracies for the baseline (B) captioning model and for the proposed approach (P) reinforced with Wikipedia geography articles.

Country	B	P	Country	B	P	Country	B	P	Country	B	P	Country	B	P
Africa														
Algeria	20	29	Angola	11	27	Benin	12	18	Botswana	4	21	Burkina Faso	14	21
Burundi	16	26	Cameroon	52	39	C. Afr. Rep.	21	26	Chad	21	27	Congo	14	21
D. R. Congo	20	29	Djibouti	16	0	Egypt	18	22	Eritrea	2	15	Ethiopia	5	21
Gabon	36	44	Ghana	27	32	Guinea	16	33	Guinea-B.	7	28	Kenya	15	22
Lesotho	26	40	Liberia	9	19	Libya	20	26	Madagascar	13	24	Malawi	25	35
Mali	18	29	Mauritania	22	55	Morocco	22	3	Mozambique	13	23	Namibia	26	30
Niger	16	27	Nigeria	23	25	Rwanda	18	18	Senegal	27	39	Sierra Leone	14	23
Somalia	25	1	Somaliland	0	0	South Africa	19	18	S. Sudan	23	34	Sudan	14	29
Tanzania	24	33	Togo	8	25	Tunisia	25	37	Uganda	13	20	Zambia	13	32
Zimbabwe	15	26												
America														
Argentina	24	34	Belize	7	7	Bolivia	16	23	Brazil	14	23	Canada	22	38
Chile	16	22	Colombia	17	29	Costa Rica	18	30	Cuba	27	37	Dominican R.	11	22
Ecuador	9	22	El Salvador	25	50	Guatemala	13	21	Guyana	0	25	Haiti	14	25
Honduras	16	32	Jamaica	0	7	Mexico	19	29	Nicaragua	22	32	Panama	30	30
Paraguay	22	32	Peru	17	29	Suriname	23	23	USA	29	45	Uruguay	0	12
Venezuela	15	0												
Asia														
Afghanistan	32	40	Armenia	22	25	Azerbaijan	17	30	Bangladesh	17	25	Bhutan	7	23
Brunei	0	33	Cambodia	19	34	China	5	19	Georgia	19	40	India	11	20
Indonesia	27	35	Iran	22	34	Iraq	18	29	Israel	27	41	Japan	16	27
Jordan	27	36	Kazakhstan	20	26	Kyrgyzstan	17	26	Laos	17	28	Lebanon	31	36
Malaysia	12	20	Mongolia	12	20	Myanmar	13	25	N. Korea	15	29	Nepal	41	46
Oman	10	25	Pakistan	24	33	Palestine	37	37	Philippines	15	26	Qatar	33	66
S. Arabia	18	24	S. Korea	16	31	Sri Lanka	22	30	Syria	34	43	Taiwan	17	31
Tajikistan	21	31	Thailand	18	30	Turkmenistan	17	22	Turkey	16	24	U.A.E.	23	35
Uzbekistan	19	26	Vietnam	13	23	Yemen	15	22						
Europe														
Albania	12	32	Armenia	22	25	Austria	26	55	Belarus	24	35	Belgium	19	34
Bosnia & H.	26	30	Bulgaria	33	42	Croatia	21	33	Czechia	24	39	Denmark	33	33
Estonia	30	41	Finland	47	51	France	21	27	Germany	33	52	Greece	25	39
Hungary	32	37	Iceland	0	16	Ireland	19	33	Italy	16	31	Kosovo	28	35
Latvia	4	11	Lithuania	24	33	Luxembourg	28	40	Moldova	25	25	Montenegro	23	37
Netherlands	11	21	N. Macedonia	10	20	Norway	15	25	Poland	24	31	Portugal	19	29
Romania	28	32	Russia	17	25	Serbia	38	44	Slovakia	23	36	Slovenia	24	30
Spain	18	51	Sweden	16	24	Switzerland	17	51	Ukraine	33	43	UK	10	16

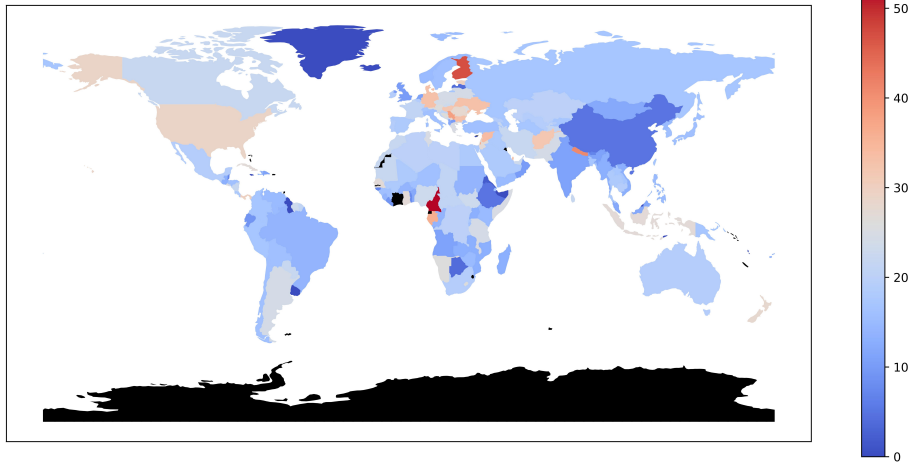


Figure 4.6 Scores of the baseline setting on each country

The main hypothesis of this thesis was that geographical domain shift can be mitigated via the inclusion of geographical metadata of the given region in the prompt of a multimodal model. To evaluate this hypothesis, we conducted extensive experiments across different architectures and compared their accuracy with and without the use of geographical context.

## 4.6 Quantitative Results

As in the SkyScript paper (Wang et al., 2024), accuracy was used as the main evaluation metric. Table 4.4 shows the performance comparison between models presented in SkyScript and our experiments.

Table 4.4 Accuracy of different models on SkyScript

Model	Accuracy
<b>SkyScript (Wang et al., 2024)</b>	
SkyCLIP-20	67.94
SkyCLIP-30	69.08
SkyCLIP-50	70.89
<b>Our Experiments</b>	
LLaMA 3.2	36.79
<b>LLaMA 3.2 + Geographical Info</b>	<b>54.95</b>
LLaVA 1.5 13B	13.05
<b>LLaVA 1.5 13B + Geographical Info</b>	<b>18.62</b>

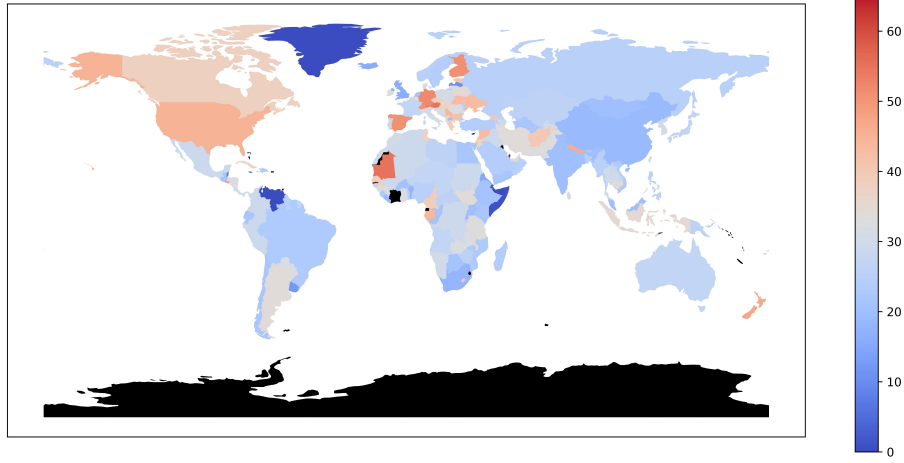


Figure 4.7 Scores of the proposed approach setting on each country

Despite the lower raw scores compared to the SkyScript baselines, the relative performance gain with the inclusion of geographical metadata is substantial—especially for the LLaMA 3.2 model, where the accuracy improved by over 18 percentage points. This consistent improvement across architectures confirms the generalizability of the proposed approach.

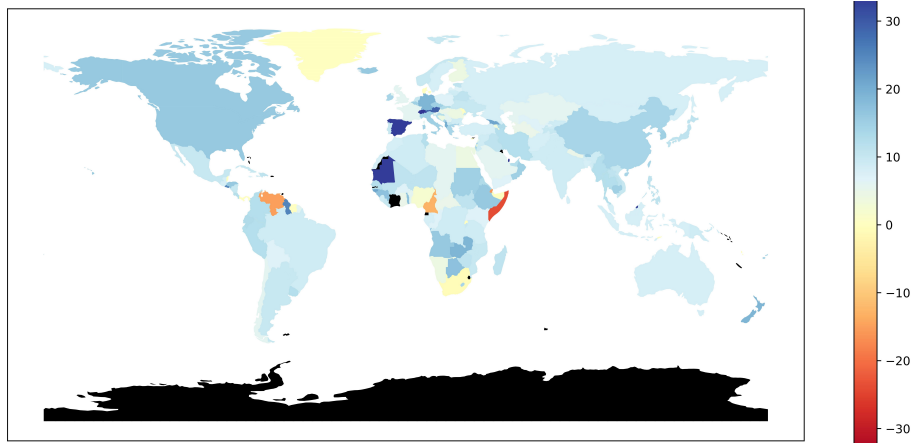


Figure 4.8 Score differences between proposed approach and baseline on each country

Examining country-level results in Figures 4.6–4.8, we observe that with the inclusion of geographical metadata, performance increased in the majority of countries. In aggregate, the accuracy improved by over 47%. Performance gains were most prominent in countries geographically closer to the European continent. On the other hand, a few countries experienced slight degradation. This may be attributed to limited or noisy geographical descriptions in Wikipedia entries, or due to heterogeneous regions within a country leading to diffuse context. Future research could explore region-specific metadata rather than national-level aggregation to address such cases.

### 4.6.1 Breakdown by Continent

To investigate whether geographic metadata benefits certain parts of the world more than others, we computed the *mean improvement* in accuracy (baseline vs. geo-informed) for each continent. The improvement for a given country was defined as

$$\Delta = \text{Accuracy}_{\text{geo}} - \text{Accuracy}_{\text{base}}.$$

We then grouped these  $\Delta$  values by continent and calculated the average.

Table 4.5 Mean accuracy improvement (%) by continent

Continent	Mean Improvement
Africa	9.0
Asia	11.2
Europe	11.4
North America	16.0
Oceania	14.0
South America	10.3

#### 4.6.1.1 Implementation Details

- We loaded the per-country scores (“scores.csv” and “geo\_scores.csv”) into a DataFrame.
- Computed `differences = geo_scores - scores`.
- Dropped any countries with missing values.
- Grouped by the `continent` column and took the mean of `differences`.

#### 4.6.1.2 Analysis

- **North America** and **Oceania** see the largest gains (16.0 and 14.0). This suggests that geographic context is especially helpful in regions whose visual appearances may diverge substantially from the European training set.

- **Europe**, the training region for all baseline models, still benefits by 11.4, indicating that even “known” domains gain from additional context.
- All continents exhibit positive improvements, confirming the general utility of country-level geographic metadata.

#### 4.6.2 Proximity to Training Region

Next, we tested whether countries “near” the training data (i.e., in Europe) benefit more than “far” countries (all others). We classified each country as:

$$\text{RegionProximity} = \begin{cases} \text{Near (Europe)} & \text{if continent} = \text{Europe}, \\ \text{Far} & \text{otherwise.} \end{cases}$$

Then we averaged the improvements within each group.

Table 4.6 Mean improvement (%) by proximity to training region

Region Proximity	Mean Improvement
Near (Europe)	11.4
Far	9.4

##### 4.6.2.1 Implementation Details

1. Added a new column “region\_proximity” based on each country’s continent.
2. Grouped by this column and computed the mean of the **differences**.

##### 4.6.2.2 Analysis

- European countries (“Near”) improve by 11.4 on average, slightly above the global mean.
- Non-European countries (“Far”) still gain 9.4 on average, demonstrating that the method generalizes beyond the immediate training domain.

- The smaller gap between “Near” and “Far” improvements ( $\approx 2$ ) indicates that geographic metadata is broadly effective, not just in regions similar to the training set.

## 4.7 Qualitative Analysis

While quantitative results demonstrate a clear performance gain through the incorporation of geographic metadata, a deeper understanding can be achieved through qualitative inspection of the generated outputs. This section highlights how access to contextual geographic cues enables models to generate more semantically appropriate and region-aware scene descriptions.

To do so, we compare predictions made by the baseline model (which has no access to geographic information) against the predictions of the geographically-informed version. Table 4.7 shows representative examples from two test countries—Norway and Kenya.

Table 4.7 Example model outputs with and without geographical information

Country	Without Geographical Info	With Geographical Info
Norway	“This image shows a large flat area with few trees under cloudy skies.”	“This image shows a Scandinavian fjord landscape, likely in Norway, with steep cliffs and overcast skies.”
Kenya	“There are hills and a river in the center of the image.”	“The image appears to be from the East African Rift Valley in Kenya, featuring semi-arid terrain and a seasonal riverbed.”

As shown, the baseline model tends to generate vague or generic descriptions that are not anchored in the geographic realities of the location. In contrast, the geographically-informed model outputs are noticeably more detailed, culturally and environmentally grounded, and specific to the country in question.

These qualitative improvements can be attributed to the integration of location metadata—such as continent, region, and Wikipedia-based textual context—which enhances the model’s understanding of the scene beyond visual features alone. This effect is especially prominent in regions with unique environmental features (e.g.,

fjords in Norway or the rift valley in Kenya), where visual appearance alone may be ambiguous or underrepresented in the pretraining data.

Moreover, this evidence supports the hypothesis that models benefit from grounding in real-world context when faced with domain shift, especially in remote sensing scenarios where visual similarity does not necessarily imply semantic similarity across regions.

### Additional Cross-Country Observations

To further explore how the model behaves across geographies, we conducted a country-level performance comparison by analyzing the *difference in scores* before and after the integration of geographic metadata. This difference indicates the relative effectiveness of our approach in different national contexts.

### Best Performing Countries

The best-performing countries are those where the addition of geographic metadata led to the largest gains in model performance. Examples include:

- **Switzerland:** Baseline score = 17, Geo-informed score = 51,  $\Delta = +34$
- **Mauritania:** Baseline score = 22, Geo-informed score = 55,  $\Delta = +33$
- **Qatar:** Baseline score = 33, Geo-informed score = 66,  $\Delta = +33$
- **Brunei:** Baseline score = 0, Geo-informed score = 33,  $\Delta = +33$
- **Spain:** Baseline score = 18, Geo-informed score = 51,  $\Delta = +33$

These large improvements highlight how the model benefits from grounding in geographically-specific context. For instance, Switzerland’s varied topography, including mountainous and alpine features, is difficult to infer solely from visual features that may resemble similar environments elsewhere. However, the inclusion of Wikipedia-derived descriptions allows the model to infer contextually relevant terms such as “Alpine”, “glacial valley”, or “Swiss plateau”.





Figure 4.9 Mauritania orthographic projection (Wikipedia contributors, 2025b).

In the case of Brunei, a country largely absent from pretraining datasets, the baseline model failed entirely, producing irrelevant or empty captions. The incorporation of metadata enabled the model to situate the image in a tropical, equatorial context, resulting in meaningful captions aligned with the region’s environmental and infrastructural characteristics.

These findings support the hypothesis that metadata can bridge representation gaps for underrepresented or visually ambiguous countries, particularly those outside the traditional centers of remote sensing datasets.

### Case Study: Mauritania

Interestingly, **Mauritania**—despite not being part of the training or validation sets—showed one of the largest performance gains after incorporating geographic metadata, with a baseline score of 22 and a geo-informed score of 55 ( $\Delta = +33$ ). This improvement cannot be attributed to overfitting or memorization and instead highlights the genuine utility of the metadata in enhancing generalization.

A close inspection of Mauritania’s Wikipedia-derived geographic description reveals a remarkably rich and detailed text. It includes specific references to prominent landforms such as the *Adrar Plateau*, the *Guelb er Richat* (also known as the “Eye of the Sahara”), and *Kediet ej Jill*, the country’s highest peak. The metadata also describes the presence of *scarps*, *sandy deserts* (ergs), and *clayey plains* (regs), along with ecologically significant terms such as *Sahel*, *Maghreb*, and *semi-arid plateaus*.

These features are not only visually distinctive in satellite imagery but also rare or underrepresented in general pretraining corpora. When made accessible to the model, such contextual clues provide strong geographical grounding. For example, references to Mauritania’s ecological zones—such as *Sahelian Acacia savanna*, *Atlantic coastal desert*, and *Saharan halophytics*—help the model differentiate it from other desert regions like Mali or Algeria, which may appear visually similar but differ ecologically.

In summary, the rich and spatially descriptive metadata allows the model to construct more accurate and semantically coherent scene captions. This case validates our hypothesis that geographic metadata can significantly enhance model performance in underrepresented regions by providing auxiliary information that complements and disambiguates visual input.

## Worst Performing Countries

On the other end of the spectrum, a few worst-performing countries experienced a *decline* in performance after incorporating geographic metadata. These include:

- **Somalia:** Baseline score = 25, Geo-informed score = 1,  $\Delta = -24$
- **Cameroon:** Baseline score = 52, Geo-informed score = 39,  $\Delta = -13$
- **Djibouti:** Baseline score = 16, Geo-informed score = 0,  $\Delta = -16$
- **Venezuela:** Baseline score = 15, Geo-informed score = 0,  $\Delta = -15$
- **Morocco:** Baseline score = 22, Geo-informed score = 3,  $\Delta = -19$

In these cases, the drop in performance may be attributed to several factors:

1. **Noisy or overly general metadata** – Wikipedia entries might emphasize political or economic narratives rather than geographical or environmental features, causing the model to misinterpret scene content.



Figure 4.10 Somalia orthographic projection (Wikipedia contributors, 2025c).

2. **Semantic drift in underrepresented regions** – In countries like *Djibouti* or *Somalia*, where pretraining data is limited and terrain may visually resemble multiple regions, the added metadata could introduce confusion instead of clarity.
3. **Mismatch between visual and textual modality** – For instance, images in Venezuela or Morocco might show urban areas or forests, while the textual metadata focuses on deserts or rural landscapes, creating inconsistency.

These results suggest that while geographic metadata often improves performance, it must be *accurate, relevant, and well-aligned* with the image context to be effective. Future research may address these shortcomings by refining metadata filtering strategies or employing learned weighting mechanisms that dynamically balance visual and textual signals.

## Case Study: Somalia

Among the countries that experienced a significant performance drop after incorporating geographic metadata, **Somalia** stands out. Its baseline score was 25, while the geo-informed score dropped to 1 ( $\Delta = -24$ ). This unexpected degradation suggests that the addition of textual metadata—rather than aiding the model—may have introduced confusion or irrelevant context.

A close examination of the Wikipedia-derived geography entry for Somalia reveals a potential reason. The entry is notably dense and multifaceted, covering a wide array of topics such as regional political divisions, historical environmental movements, toxic waste scandals, seasonal winds, and international marine law. While the metadata does contain relevant physical geography—such as descriptions of the *Ogo Mountains*, *Guban coastal plains*, and *Shabelle and Jubba rivers*—these insights are buried amidst socio-political and ecological commentary that may overwhelm or distract the model.

Another issue is that while Somalia’s terrain is geographically diverse (ranging from semi-arid plateaus to coastal mangroves), much of the descriptive language is abstract or general. Phrases such as “hot conditions prevail year-round,” “periodic monsoon winds,” or “region shaped like a tilted number seven” may be unhelpful in grounding visual content. The mention of radiation sickness and illegal toxic waste disposal—though significant in a real-world context—is likely irrelevant or even detrimental for a model learning visual scene classification.

In contrast to Mauritania’s concise and spatially structured metadata, Somalia’s entry lacks a coherent narrative aligned with visually discriminative geographic features. The inclusion of verbose political and environmental activism content may dilute the signal-to-noise ratio, limiting the model’s ability to extract useful geographic priors for scene captioning.

This case illustrates the importance of metadata quality, relevance, and alignment with the visual domain. It also raises a need for filtering or structuring metadata to focus on geographic and ecological content directly associated with the scene-level appearance of the region.

## 5. Conclusion and Future Work

We have proposed an approach to tackle domain adaptation using multimodal models with geographic metadata. Providing multimodal models with geographic metadata from the image boosts performance and reduces the impact of domain shift in remote sensing image captioning. As far as we know, this is the first implementation of Wikipedia geographic data to reduce geographical domain shift. In future work, this implementation can be extended to encompass additional computer vision tasks, including object detection, and visual question answering.

As for future work, this study can be extended by incorporating **regional-level geographic information** instead of relying solely on **country-level metadata**. While using country-level geography has shown promising results, it introduces a notable limitation: many countries encompass a wide variety of geographic features, climates, and landscapes within their borders. As a result, a single, aggregated geographic description often fails to capture the internal diversity present within a nation.

For instance, the Central Anatolia region in Türkiye—illustrated in Figure 5.1—represents just one of several distinct geographic regions within the country, each characterized by different topographical, ecological, and cultural attributes.



Figure 5.1 Central Anatolia Region (Wikipedia contributors, 2025a).

Adopting a **region-specific approach** would allow for a more nuanced understanding of image content by leveraging geographically localized context. This refinement could significantly enhance classification accuracy and semantic alignment, particularly in scenarios where the diversity within a country leads to visual and contextual ambiguities. In other words, geographic metadata tailored to specific regions could offer richer and more precise cues, helping the model disambiguate between visually similar scenes that occur in vastly different environments.

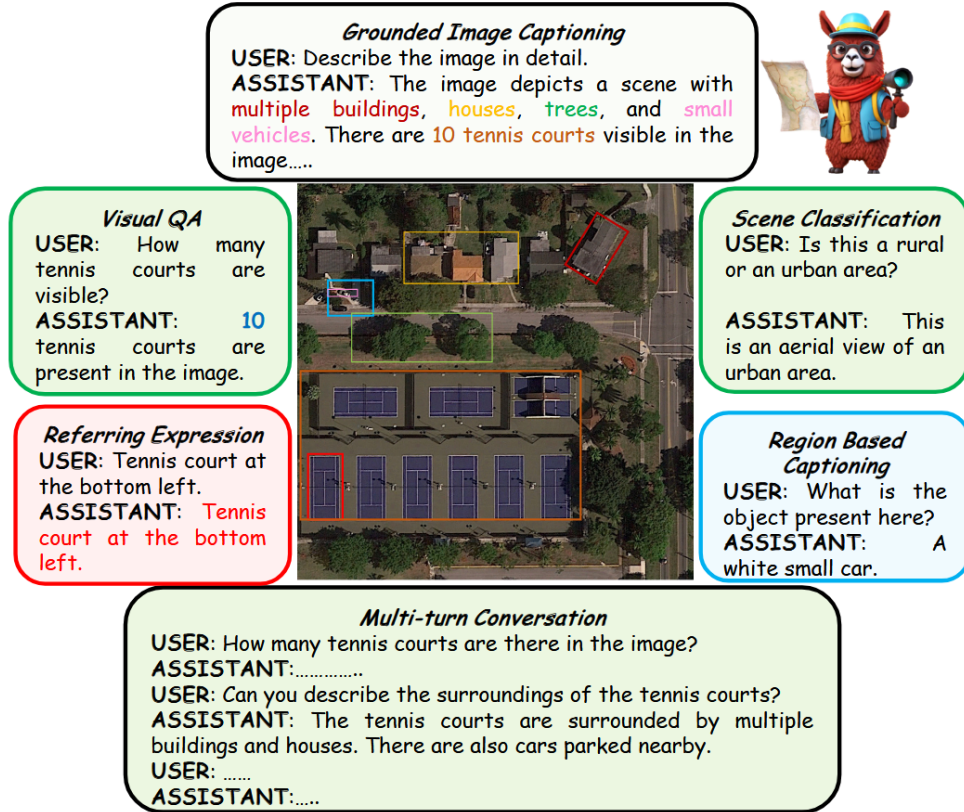


Figure 5.2 GeoChat tasks (Kuckreja et al., 2024).

In addition to regional granularity, another promising direction for future research is to **evaluate the generalizability of this approach across various computer vision tasks beyond scene classification**. The methodology proposed in this thesis—enriching visual understanding through geographic grounding—can potentially be applied to tasks such as *object detection*, *image captioning*, *visual question answering*, or *domain adaptation*. Each of these tasks stands to benefit from a deeper understanding of how geographic context influences semantic content. Recognizing and accounting for regional or domain-specific shifts could lead to performance gains and increased model robustness, particularly in real-world deployment scenarios where such shifts are prevalent. Fig 5.2 shows multiple visual tasks that Geochat model can do.

## Ablation Study on Metadata Components

While this thesis demonstrates that incorporating Wikipedia-based geographic metadata enhances performance, future work can isolate which components of that metadata are most effective by conducting an **ablation study**. Specifically, different subsets of the metadata can be tested independently to determine their individual contributions. We propose the following ablation variants:

- **Only the Country Name:** Include only the name of the country (e.g., This image is from Kenya) without any additional descriptive metadata. This tests whether geographic anchoring alone is sufficient to guide the model, possibly through memorized prior associations.
- **Only Climate or Land Use Information:** Include descriptions of climate zones (e.g., arid steppe, humid subtropical) and common land cover types (e.g., agricultural plains, forested hills), without naming the country. This tests whether purely environmental cues—closely aligned with visual appearance—drive the improvement.
- **Only Population/Density Descriptions:** Include demographic indicators such as population size, density, and urbanization (e.g., densely populated urban areas along the coast, with sparsely inhabited interior). This tests whether human geography alone provides discriminative power for classification.

Each of these subsets can be tested using the same model architecture and training procedure used in the current study, modifying only the metadata content in the prompt. Classification accuracy from each ablation setting can then be compared to the full metadata setting and the no-metadata baseline.

Listing 5.1 Example JSON conversation data

```
{
  "image": "images6/n1567088198_S2_18.jpg",
  "conversations": [
    {
      "from": "human",
      "value": "<image>\nThis image is taken in {country_name
        }. Classify the image with a single label from [...]
        You must provide a single label."
    },
    {
      "from": "gpt",
      "value": "place of village"
    }
  ]
}
```

Such an ablation study would offer the following benefits:

- **Interpretability:** Clarifies which types of metadata are most semantically valuable to the model.
- **Efficiency:** Helps streamline the amount of metadata needed, reducing complexity in real-time systems.
- **Robustness:** Evaluates whether the model is overly reliant on specific metadata types or performs well across various information sources.

This analysis would provide a more nuanced understanding of why geographic metadata works and help guide future development of more efficient and robust multi-modal pipelines for remote sensing applications.



## BIBLIOGRAPHY

- Aljanaki, A., Yang, Y.-H., and Soleymani, M. (2017). Developing a benchmark for emotional analysis of music. *PLOS ONE*, 12(3):e0173392.
- Balaji, Y., Sankaranarayanan, S., and Chellappa, R. (2018). Metareg: Towards domain generalization using meta-regularization. In *Advances in Neural Information Processing Systems*, volume 31, pages 1006–1016, Montréal, Canada. Curran Associates, Inc.
- Bazi, Y., Bashmal, L., Al Rahhal, M. M., Ricci, R., and Melgani, F. (2024). Rs-llava: A large vision-language model for joint captioning and question answering in remote sensing imagery. *Remote Sensing*, 16(9):1477.
- Benjdira, B., Bazi, Y., Koubaa, A., and Ouni, K. (2019). Unsupervised domain adaptation using generative adversarial networks for semantic segmentation of aerial images. *Remote Sensing*, 11(11):1369.
- Camgoz, N. C., Hadfield, S., Koller, O., Ney, H., and Bowden, R. (2018). Neural sign language translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7784–7793.
- Cha, K., Yu, D., and Seo, J. (2024). Pushing the limits of vision-language models in remote sensing without human annotations. *arXiv preprint arXiv:2409.07048*.
- Chen, K., Liu, C., Chen, H., Zhang, H., Li, W., Zou, Z., and Shi, Z. (2024). Rsprompter: Learning to prompt for remote sensing instance segmentation based on visual foundation model. *IEEE Transactions on Geoscience and Remote Sensing*, 62.
- Dai, W., Li, J., Li, D., Tiong, A. M. H., Zhao, J., Wang, W., Li, B., Fung, P. N., and Hoi, S. (2024). Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36.
- DeepLearning.AI (2025a). Attention in transformers: Concepts and code in pytorch. <https://learn.deeplearning.ai/courses/attention-in-transformers-concepts-and-code-in-pytorch/lesson/han2t/introduction>.
- DeepLearning.AI (2025b). Finetuning large language models. <https://www.deeplearning.ai/short-courses/finetuning-large-language-models/>.
- DeepLearning.AI (2025c). How diffusion models work. <https://www.deeplearning.ai/short-courses/how-diffusion-models-work/>.
- DeepLearning.AI (2025d). Prompt engineering for vision models. <https://www.deeplearning.ai/short-courses/prompt-engineering-for-vision-models/>.

- Dou, Q., de Castro, D. C., Kamnitsas, K., and Glocker, B. (2019). Domain generalization via model-agnostic learning of semantic features. *arXiv preprint arXiv:1910.13580*.
- Fang, L., Kuang, Y., Liu, Q., Yang, Y., and Yue, J. (2023). Rethinking remote sensing pretrained model: Instance-aware visual prompting for remote sensing scene classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–13.
- French, G., Mackiewicz, M., and Fisher, M. (2018). Self-ensembling for visual domain adaptation. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Ganin, Y. and Lempitsky, V. (2016). Domain-adversarial training of neural networks. *The Journal of Machine Learning Research*, 17(1):2096–2030.
- Getty Images (2025). Two different domains. <https://www.gettyimages.com/>.
- Hoffman, J., Tzeng, E., Park, T., Zhu, J.-Y., Isola, P., Saenko, K., Efros, A. A., and Darrell, T. (2018). Cycada: Cycle-consistent adversarial domain adaptation. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1989–1998. PMLR.
- Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2022). Lora: Low-rank adaptation of large language models. In *Proceedings of the 10th International Conference on Learning Representations (ICLR)*, pages 1–15. OpenReview.net.
- Hu, Y., Yuan, J., Wen, C., Lu, X., and Li, X. (2023). Rsgpt: A remote sensing vision language model and benchmark. *arXiv preprint arXiv:2307.15266*.
- Jalammar, J. (2016). The illustrated word2vec. <https://jalammar.github.io/illustrated-word2vec/>.
- Jiang, H., Yin, J., Wang, Q., Feng, J., and Chen, G. (2025). Eaglevision: Object-level attribute multimodal llm for remote sensing. *arXiv preprint arXiv:2503.23330*.
- Kuckreja, K., Danish, M. S., Naseer, M., Das, A., Khan, S., and Khan, F. S. (2024). Geochat: Grounded large vision-language model for remote sensing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27831–27840.
- Lee, D.-H. (2013). Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *Proceedings of the ICML 2013 Workshop on Challenges in Representation Learning*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 896–904, Atlanta, GA, USA. Awarded second prize in the Black Box Learning Challenge.
- Li, Y., Wang, N., Shi, J., Liu, J., and Hou, X. (2016). Revisiting batch normalization for practical domain adaptation. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

- Liu, C., Zhao, R., Chen, J., Qi, Z., Zou, Z., and Shi, Z. (2023). A decoupling paradigm with prompt learning for remote sensing image change captioning. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–13.
- Liu, F., Chen, D., Guan, Z., Zhou, X., Zhu, J., Ye, Q., Fu, L., and Zhou, J. (2024a). Remoteclip: A vision-language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62.
- Liu, H., Li, C., Li, Y., and Lee, Y. J. (2024b). Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26296–26306.
- Liu, M.-Y. and Tuzel, O. (2016). Coupled generative adversarial networks. In *Advances in Neural Information Processing Systems*, pages 469–477.
- Liu, Y., Zhang, K., Li, Y., Yan, Z., Gao, C., Chen, R., Yuan, Z., Huang, Y., Sun, H., Gao, J., et al. (2024c). Sora: A review on background, technology, limitations, and opportunities of large vision models.
- Lobry, S., Marcos, D., Murray, J., and Tuia, D. (2020). Rsvqa: Visual question answering for remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 58(12):8555–8566.
- Long, M., Cao, Y., Wang, J., and Jordan, M. I. (2015). Learning transferable features with deep adaptation networks. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 97–105. PMLR.
- Mall, U., Phoo, C. P., Liu, M. K., Vondrick, C., Hariharan, B., and Bala, K. (2023). Remote sensing vision-language foundation models without annotations via ground remote alignment. *arXiv preprint arXiv:2312.06960*.
- Mehmood, M., Shahzad, A., Zafar, B., Shabbir, A., and Ali, N. (2022). Remote sensing image classification: A comprehensive review and applications. *Mathematical Problems in Engineering*, 2022(1):5880959.
- Meta AI (2024). Llama 3.2 connect 2024: Vision for edge and mobile devices. <https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/>.
- Nedungadi, V., Kariryaa, A., Oehmcke, S., Belongie, S., Igel, C., and Lang, N. (2024). Mmearth: Exploring multi-modal pretext tasks for geospatial representation learning. *arXiv preprint arXiv:2405.02771*.
- Patel, V. M., Gopalan, R., Li, R., and Chellappa, R. (2015). Visual domain adaptation: A survey of recent advances. *IEEE Signal Processing Magazine*, 32(3):53–69.
- Pinecone (2025). A beginner’s guide to vector embeddings. <https://www.pinecone.io/learn/vector-embeddings/>.
- Plummer, B. A., Wang, L., Cervantes, C. M., Caicedo, J. C., Hockenmaier, J., and Lazebnik, S. (2015). Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2641–2649.

- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 8748–8763.
- Roberts, J., Lüddecke, T., Sheikh, R., Han, K., and Albanie, S. (2024). Charting new territories: Exploring the geographic and geospatial capabilities of multi-modal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 554–563. IEEE.
- Saha, S., Zhao, S., and Zhu, X. X. (2022). Multitarget domain adaptation for remote sensing classification using graph neural network. *IEEE Geoscience and Remote Sensing Letters*, 19:6506505.
- Sharma, P., Ding, N., Goodman, S., and Soricut, R. (2018). Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.
- Silva, J. D., Magalhães, J., Tuia, D., and Martins, B. (2024). Multilingual vision-language pre-training for the remote sensing domain. In *Proceedings of the 32nd ACM International Conference on Advances in Geographic Information Systems*, pages 220–232.
- Sun, B. and Saenko, K. (2016). Deep coral: Correlation alignment for deep domain adaptation. In *European Conference on Computer Vision (ECCV) Workshops*, pages 443–450. Springer.
- Tian, P., Yang, Y., and Wei, Y. (2024). Domain alignment with large vision-language models for cross-domain remote sensing image retrieval. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM)*, pages 1234–1245. ACM.
- Tuia, D., Persello, C., and Bruzzone, L. (2016). Domain adaptation for the classification of remote sensing data: An overview of recent advances. *IEEE Geoscience and Remote Sensing Magazine*, 4(2):41–57.
- Tzeng, E., Hoffman, J., Saenko, K., and Darrell, T. (2017). Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7167–7176.
- Tzeng, E., Hoffman, J., Zhang, N., Saenko, K., and Darrell, T. (2014). Deep domain confusion: Maximizing for domain invariance. arXiv preprint arXiv:1412.3474.
- Vidit, V., Engilberge, M., and Salzmann, M. (2023). Clip the gap: A single domain generalization approach for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3219–3229.
- Wang, M. and Deng, W. (2018). Deep domain adaptation: A survey. *Neurocomputing*, 312:135–153.

- Wang, Z., Prabha, R., Huang, T., Wu, J., and Rajagopal, R. (2024). Skyscript: A large and semantically diverse vision-language dataset for remote sensing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 5805–5813.
- Wikipedia contributors (2025a). Central anatolia region. [https://en.wikipedia.org/wiki/Central\\_Anatolia\\_region](https://en.wikipedia.org/wiki/Central_Anatolia_region).
- Wikipedia contributors (2025b). Mauritania. <https://en.wikipedia.org/wiki/Mauritania>.
- Wikipedia contributors (2025c). Somalia. <https://en.wikipedia.org/wiki/Somalia>.
- Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. (2014). From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.
- Yuan, J., Hou, F., Du, Y., Shi, Z., Geng, X., Fan, J., and Rui, Y. (2022). Self-supervised graph neural network for multi-source domain adaptation. In *Proceedings of the 30th ACM International Conference on Multimedia (ACM MM 2022)*, pages 3907–3916.
- Yuan, Z., Xiong, Z., Mou, L., and Zhu, X. X. (2024). Chatearthnet: A global-scale image-text dataset empowering vision-language geo-foundation models. arXiv preprint arXiv:2402.11325.
- Zhan, Y., Xiong, Z., and Yuan, Y. (2025). Skyeyegpt: Unifying remote sensing vision-language tasks via instruction tuning with large language model. *ISPRS Journal of Photogrammetry and Remote Sensing*, 221:64–77.
- Zhang, Y., Zhang, M., Li, W., Wang, S., and Tao, R. (2023). Language-aware domain generalization network for cross-scene hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–12.
- Zhang, Z., Zhao, T., Guo, Y., and Yin, J. (2024). Rs5m and georsclip: A large scale vision-language dataset and a large vision-language model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*.
- Zhao, Q., Lyu, S., Zhao, H., Liu, B., Chen, L., and Cheng, G. (2024). Self-training guided disentangled adaptation for cross-domain remote sensing image semantic segmentation. *International Journal of Applied Earth Observation and Geoinformation*, 127:103646.
- Zheng, Z., Lv, L., He, J., and Zhang, L. (2025). Unirs: Towards unified multi-task fine-tuning for remote sensing foundation model. *IEEE Transactions on Geoscience and Remote Sensing*, 63(5):1234–1245.
- Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A. (2017). Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2223–2232.

## APPENDIX A

Kerem Aydın, Erchan Aptoula. Domain Generalized Remote Sensing Scene Captioning via Country-Level Geographic Information. *In Proceedings of the Joint Urban Remote Sensing Event (JURSE)*, 2025.