Generative and Retrieval Image Captioning Towards Automated InSAR Image Analysis

Joe Yazbeck * and John B. Rundle

Posted Date: 12 May 2026

doi: 10.20944/preprints202605.0685.v1

Keywords: natural language processing; machine learning; InSAR; image captioning; highperformance computing; remote sensing; vision-language models; transformer models; volcanic deformation; seismology

 doi:10.20944/preprints202605.0685.v1

Article

Generative and Retrieval Image Captioning Towards Automated InSAR Image Analysis

Joe Yazbeck [1,*] and John B. Rundle [1,2,3]

[1] Department of Physics and Astronomy, University of California, Davis, Davis, CA, 95616, USA
[2] Department of Earth and Planetary Sciences, University of California, Davis, Davis, CA, 95616, USA
[3] Santa Fe Institute, Santa Fe, NM 87501, USA
* Correspondence: jyazbeck@ucdavis.edu

Simple Summary

Satellite radar images are widely used to monitor ground deformation related to earthquake and volcanic activity. However, interpreting these images usually requires specialized expertise and careful manual inspection. As the volume of satellite data continues to grow, analyzing these images quickly and consistently becomes increasingly challenging. In this study, we investigate whether modern artificial intelligence models can automatically produce written descriptions of satellite radar images. By training computer vision and language models on previously labeled examples, the system learns to associate visual patterns with meaningful textual explanations. The results suggest that these models can capture important features of deformation and generate useful descriptions of the images. This work represents a first step toward automated interpretation tools for large-scale geophysical monitoring datasets.

Abstract

Interpreting interferometric synthetic aperture radar (InSAR) imagery is a critical task in monitoring volcanic and seismic activity, yet the process usually requires expert knowledge and manual analysis. As the volume of satellite observations continues to increase, automated methods capable of describing and interpreting these images become increasingly important in order to assist geophysical monitoring efforts. In this work, we investigate the feasibility of automated image captioning for InSAR data using modern vision-language models. We utilize the Hephaestus dataset which is a large collection of annotated interferograms focused on volcanic deformation, and apply a series of preprocessing steps to curate a balanced dataset of deforming and non-deforming images. Two generative image captioning architectures, the Generative Image-to-Text Transformer (GIT) and Bootstrapping Language-Image Pretraining (BLIP), are fine-tuned to output natural language descriptions of the InSAR images. In addition, we implement a retrieval-based model that aligns image and text representations within a shared embedding space and retrieves the most semantically similar caption. The performance of these approaches is evaluated using standard captioning metrics and qualitative inspection of generated descriptions. Our results suggest that pre-trained vision–language models can adapt to specialized scientific imagery despite being trained primarily on natural image datasets. This study represents an initial step towards automated interpretation systems capable of assisting researchers in large-scale InSAR monitoring applications.

Keywords: natural language processing; machine learning; InSAR; image captioning; high-performance computing; remote sensing; vision-language models; transformer models; volcanic deformation; seismology

1. Introduction

Image captioning refers to the task of translating visual information into meaningful text [1]. While it is, in general, fairly simple for a human to look at an image, extract the key features, and form

 doi:10.20944/preprints202605.0685.v1

a detailed description of the given image, this endeavor becomes a much more complicated exercise when assigned to machines [2]. In fact, this task has long been viewed as a challenging problem lying at the very intersection of computer vision and natural language processing [3].

Early efforts to tackle this problem generally focused on algorithms that would patch a given static image in order to detect different objects and map each object to a corresponding descriptive word [4,5]. This was implemented using a variety of different techniques [6,7]. For example, Farhadi et al. defined a meaning space in their approach that linked the space of images to the space of sentences [8]. By evaluating the similarity score between an image and a sentence in the meaning space, they were able to assign appropriate captions based on various error metrics [8].

These methods steadily shifted into using templates to generate image descriptions [9]. To illustrate, Kulkarni et al. implemented linguistic constraints through the use of simple templates such as: "This is a photograph of ..." or "Here we see ..." [10]. This was done in order to force sentences to come out grammatically correct and structurally coherent at a time when language models alone could not reliably produce this [10]. Despite the fact that this methodology inhibits the generation of novel captions, the results showed great promise in developing a system to automatically generate natural language captions for images based on satisfactory ROUGE and BLEU scores, the evaluation metrics commonly used within the natural language processing community [10–12].

Template-based techniques were further improved upon by Mitchell et al. [13]. To illustrate, rather than relying on a fixed template to generate captions, Mitchell et al. took a different approach by generating syntactic trees from vision detections [13]. Their system, called 'Midge', leveraged word co-occurrence statistics to detail what the computer vision system sees and produced natural captions that significantly outperformed previous automatic systems [13].

It wasn't until researchers began to incorporate deep neural networks into their models that image captioning performance truly and rapidly increased [14–17]. This was inspired in large part by the groundbreaking results of the deep convolutional neural network AlexNet on the ImageNet database [18]. The first methods framed the problem as a ranking problem and employed a retrieval-based approach. To illustrate, Socher et al. introduced a model that would learn to map images and sentences into a shared embedding space using a deep neural network [19]. This would then enable the retrieval of one given the other [19]. A pairwise ranking loss function is minimized during training to learn to rank images and their descriptions [19]. Their model outperformed baselines as well as other commonly used models [19].

The main issue with this approach is the inability to generate novel descriptions or visual depictions from the embedding space. In other words, it was not feasible to perform the inverse projection. This is why Chen et al. sought to explore the bi-directional mapping between images and their descriptions using recurrent neural networks [20]. The key contribution was the addition of a hidden layer to the visual recurrent network that would allow the model to remember visual concepts over the long term [20]. This allows the network to automatically choose the relevant concepts that it trained on in order to describe a given image [20]. State-of-the-art results were achieved based on measurements of BLEU, METEOR, and CIDEr scores [11,20–22].

Additional work aimed at simplifying the traditional pipeline using advances in machine translation. In fact, Vinyals et al. formulated an end-to-end system that combines the processes of object detection and caption generation into a single joint model [23]. The idea was to take an image input and train the model to maximize the likelihood of generating a specific target word sequence that is pulled from a given dictionary [23]. The images are first passed through a CNN-based encoder that transforms them into meaningful vector representations which in turn are used as the initial hidden state of the RNN-based decoder that would generate the target sentence [23]. By applying the principles of machine translation for generating descriptions, they were able to achieve state-of-the-art results on standardized datasets based on the BLEU-1 and BLEU-4 metrics [23].

A major performance improvement emerged with the incorporation of the transformer architecture, which was first introduced by Vaswani et al. for the purposes of machine translation and

 doi:10.20944/preprints202605.0685.v1

language understanding tasks [24]. The architecture's key feature is the self-attention mechanism that essentially connects one element of a set to all the others using a scaled dot product [24].

This approach, along with its variants, dominated the natural language processing field and later the computer vision field as well [1]. In fact, one of the first image captioning models to incorporate self-attention used a module to encode the relationships between features resulting from an object detector [25]. The full transformer architecture was first investigated by Li et al. in the realm of image captioning and sequence modeling [26]. Along with their introduction of entangled attention, they were able to achieve state-of-the-art performance on the MS COCO dataset, negating the need for RNNs as their complexity can be a disadvantage [26].

Image captioning models saw many additional modifications and variants to the attention mechanism following the introduction of the transformer architecture. To start with, Pan et al. introduced a unified attention block that uses bilinear pooling to highlight attended visual features [27]. This technique led to an enhanced set of image-level and region-level features due to the higher order intra-modal interactions [27]. Moreover, Cornia et al. proposed a memory-augmented attention operator by adding additional slots to the set of keys and values used for self-attention in order to encode a priori information [28]. A meshed connectivity between the encoding and decoding modules allowed them to achieve a new state-of-the-art performance on COCO based on standard metrics [28].

Inspired by the introduction of the vision transformer (ViT) [29], Liu et al. developed the first convolution-free image captioning model [30]. To elaborate, their architecture used a pre-trained ViT as an encoder and a standard transformer as a decoder to generate captions. This same approach was later used to train large visual encoder models such as CLIP [31] and SimVLM [32] that are capable of performing many downstream tasks including image captioning.

The importance of tackling this problem is significant as advances in image captioning can directly benefit society. Applications include, but are not limited to, aiding in navigation for visually impaired people, improving early childhood education, and enabling improved computer-human interactions, to name a few [33,34]. This sets a precedent for researchers to explore these techniques in a range of disciplines that are not specifically related to natural images only [35,36]. For example, Gamidi et al. applied image captioning techniques to describe underwater scenery where visual conditions generally make it difficult to conduct marine research and environmental monitoring [37]. Since most image captioning datasets are of natural and terrestrial scenes, they resorted to applying underwater-specific augmentations to these images before passing them to training in their transformer-based image captioning model [37]. This allowed the model to detect underwater features and perform well on the test set during inference [37].

Due to rapid advances in computer science, engineering, and aerospace systems, modern remote sensing technologies are able to provide researchers with high-resolution satellite images [38]. This encouraged the use of image captioning in the remote sensing community, especially since the applications are wide, such as describing photos captured by unmanned aerial vehicles (UAVs) in war, reconnaissance, and rescue operations [39]. Moreover, UAVs are generally able to quickly access rugged terrains allowing them to analyze and pinpoint the needy in natural disaster situations such as earthquakes, floods, and volcanic eruptions [40]. In fact, Qu et al. were one of the first to bridge the gap between these two fields by applying a standard CNN along with a LSTM/RNN to extract image features and combine them with text descriptions [41]. This resulted in good semantic understanding of high-resolution remote sensing images [41].

Remote sensing has proven to be a powerful and essential tool in monitoring and assessing earthquake and volcanic hazards [42,43]. Forces acting deep within Earth's interior result in measurable deformation on the surface providing the necessary data to improve numerical models for forecasting these catastrophic events [44]. Specifically, the interferometric synthetic aperture radar (InSAR) technique excels in this domain as it is able to directly measure the change in height of the ground along the line-of-sight of the satellite or plane [45]. For example, Biggs et al. conducted an InSAR study on the 1994 and 2004 Al Hoceima earthquakes and were able to modify previous tectonic

 doi:10.20944/preprints202605.0685.v1

model predictions on the fault plane orientation providing key insights on the tectonics of that area in Morocco [46]. Additionally, InSAR has been extensively used to monitor similar phenomena dealing with land subsidence and uplift making it a reliable tool for hazard and risk assessment [47,48]. In volcanology, InSAR has been used to detect surface deformation which has a strong statistical link to eruption [49]. Bountos et al. used a vision transformer model and trained it on synthetically generated interferograms to detect volcanic unrest, and they were able to achieve state-of-the-art accuracies on the test set [50].

In general, interpreting InSAR images can be a challenging task for non-experts as there are many complexities involved with the processing steps as well as the resulting errors making the need for expert input critical [51–54]. Additionally, the recent developments in technology and computational power have resulted in large unprecedented quantities of monitoring data such as InSAR which cannot be realistically inspected manually anymore [49]. This highlights the need to have an automated InSAR image captioning system for rapid analysis that would be extremely beneficial especially in cases such as immediately following or preceding catastrophes. To the best of our knowledge, this task has yet to be performed.

In this paper, we make the first attempt at achieving an automated InSAR image captioning system. We explore two different techniques in image captioning applied to the problem of InSAR labeling using the Hephaestus dataset. Specifically, we evaluate two generative captioning architectures called GIT and BLIP that are fine-tuned on the domain-specific imagery, and we compare their performance to a custom-built retrieval-based model that aligns visual and textual representations within a shared embedding space. By analyzing both the generative and retrieval paradigms, we seek to assess the feasibility of utilizing vision-language models in specialized remote sensing data. Our results provide insight into the strengths and limitations of these approaches and highlight the potential for automated image captioning systems to assist large-scale geophysical monitoring workflows.

2. Materials and Methods

2.1. Hephaestus Dataset

An ode to the Greek God of fire and volcanoes, the Hephaestus dataset created by Bountos et al. consists of annotated InSAR images focused on global volcano monitoring [55]. Specifically, data retrieved from the Comet-LiCS portal monitors 44 of the most active volcanoes globally over the period of 2014 to 2021 [56–58]. The wrapped InSAR images correspond to 38 different ascending and descending Sentinel-1 frames since, in some cases, more than one volcano could be included in one frame.

The dataset was manually annotated by experts in the field who leveraged not only their InSAR interpretation skills, but also the extensive literature regarding the historical activities of the volcanoes. This is primarily due to the difficulty of separating atmospheric effects from actual deformation, as they look very similar in an InSAR image. The result is a dataset that contains 19,919 samples, of which 1,833 contain ground deformation as a result of volcanic activity or an earthquake.

Each interferogram was annotated with 20 different label categories. The first few labels refer to the metadata of the image, such as the frame ID, which gives the satellite orbit and frame, and the primary and secondary dates, which give the acquisition dates of the SAR image.

The next class of labels describes any technical errors that may have occurred in the automatic InSAR processing and generation by Comet-LiCS. A binary category separates corrupted interferograms that are totally problematic from usable ones. Another category highlights the different types of processing errors that may be present, such as debursting or merging errors.

The presence of various kinds of fringes is also noted in each image's annotation. For example, glacier fringes caused by glacier melting as well as orbital fringes caused by a phase ramp due to orbital errors are included in the annotation. Moreover, fringes caused by different atmospheric distortions and effects, such as the varying refractive index of the troposphere and the vapors caused by liquid and solid particles of the atmosphere, are also noted. Additionally, the annotation includes information

 doi:10.20944/preprints202605.0685.v1

on whether the image has low coherence due to interferometric signal decorrelation. A label is also included to indicate whether there are artifacts on the image itself such as an artificial colorbar legend. The latter is helpful in understanding the significance of the color in each image.

The last set of labels describes the actual deformation, if present, in the image. A classification of the volcanic deformation activity according to the magma source is also provided. This activity type will have a value of Mogi, Dyke, Sill, Spheroid, or Unidentified. The Mogi model represents the underground magma chamber as a point source of pressure and typically produces a radially symmetric pattern of subsidence or uplift [59,60]. A dyke is a vertical sheet of magma that cuts through existing rock layers and generally creates a linear deformation pattern with subsidence and uplift on either side of the dyke [61,62]. A sill is a horizontal sheet of magma that intrudes between existing rock layers [63]. The resulting surface deformation is a broad and gentle uplift that is circular, but less pronounced than a Mogi source [64]. A spheroid models the magma source as a finite-sized spheroid or ellipsoid, allowing for an elongated shape in 3D space [65]. This produces a radially symmetric uplift that may be slightly elongated depending on the aspect ratio of the spheroid. An important label included in the annotation is the intensity of the observed deformation. Additional labels indicating the phase of the deformation as well as the confidence of the annotation itself are also highlighted. A manually-drawn segmentation mask specifying the displacement area is provided. Finally, the last label is a full description of the interferogram in the form of a caption that cohesively ties most of the previous labels together.

The dataset itself is intentionally designed to be highly diverse in order to capture a wide range of observational and processing conditions. For instance, images from both ascending and descending viewing geometries are used along with multiple range-azimuths looks combinations. Additionally, different color scale palettes can be found within.

A pronounced class imbalance is exhibited between samples showing deformation and those showing no deformation. This is to be expected as volcanic activity and seismic events are relatively rare phenomena to occur in nature. Among deformation samples, further imbalances exist across deformation types, event intensities, and volcanic phases with most events corresponding to unrest phases and classified as sill types. An example deforming image from the Hephaestus dataset is shown in Figure 1 along with its corresponding caption (unique ID: 1813). An example non-deforming image is shown in Figure 2 (unique ID: 15281). A color fringe in the direction of blue-green-yellow-orange-purple-blue means a change of 2π radians towards the satellite [56].

 doi:10.20944/preprints202605.0685.v1

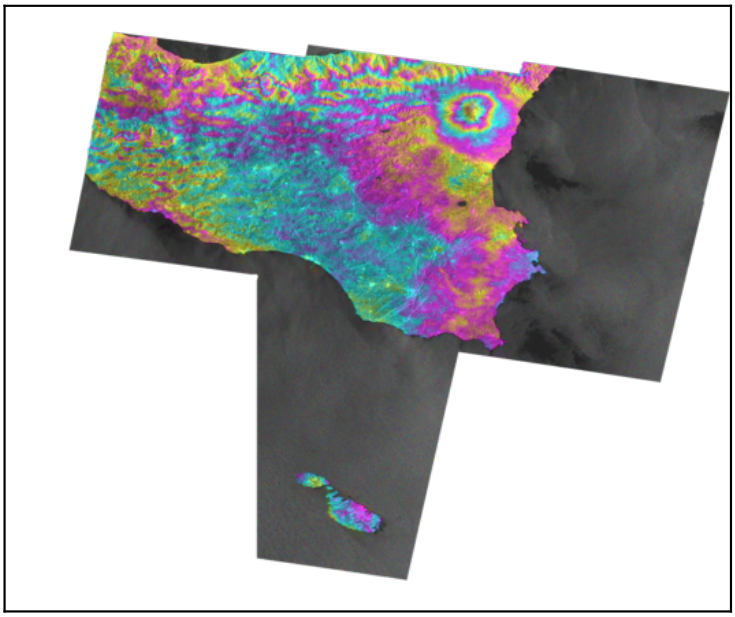

Figure 1. Vertical stratification effect can be detected on the top side of the region. Wave-like patterns caused by atmospheric delays can also be detected on the top and left side. A mogi-type deformation pattern of low intensity is detected in top-right side of the region.

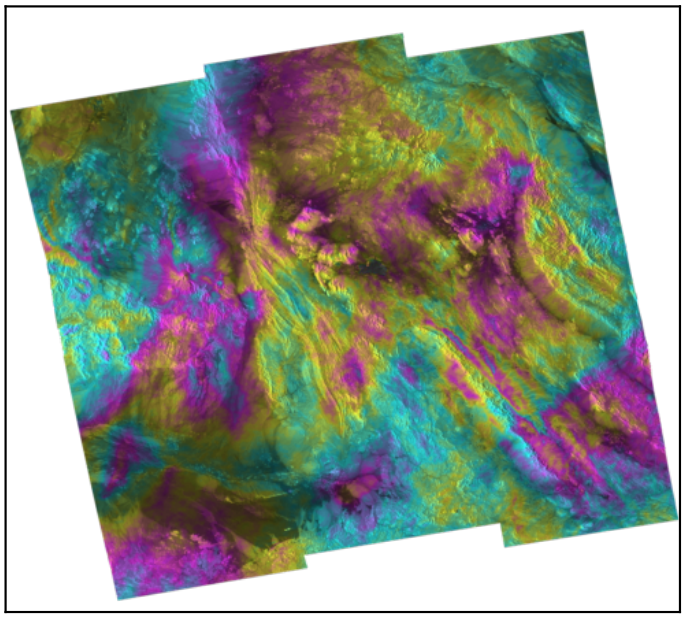

Figure 2. Turbulent mixing effect or wave-like patterns caused by liquid and solid particles of the atmosphere can be detected around the area. No deformation activity can be detected.

 doi:10.20944/preprints202605.0685.v1

2.2. Preprocessing

Arguably one of the most critical steps in any machine learning framework, dataset preprocessing directly influences a model's learned representations and predictive behavior. This is especially true for multi-modal tasks involving scientific imagery where slight inconsistencies in image formats or visual artifacts can introduce undesired biases [66]. To mitigate these effects, we applied a series of preprocessing and curation steps to the Hephaestus dataset while mainly focusing on image quality, format consistency, and class balance.

In order to ensure uniformity when it came to the quality of images used, we selectively chose certain conditions to be met by the images according to the labels in each of their respective annotation. To start with, images with any value besides 0 in the labels of *corrupted*, *no_info*, *processing_error*, or *low_coherence* were discarded. This immediately eliminates many images that are simply unusable or very difficult to extract information from. Similarly, we ignored images containing any value other than 0 in the labels of *orbital_fringes* and *glacier_fringes* as we did not want any fringes related to orbital errors or fringes related to glaciers melting as that is not the scope of this paper. Images with *atmospheric_fringes* were kept as long as that value was less than or equal to 2. Fringes related to atmospheric effects are to be expected and, in fact, most images in the dataset exhibited some form of atmospheric fringes. The values of this label ranged from 0 to 3. A value of 0 indicated that there was little to no atmospheric effect. Values 1 and 2 indicated changes related to the troposphere's refractive index and changes related to the turbulent mixing and vapors of solid and liquid particles in the atmosphere, respectively. A value of 3 indicated the presence of both effects 1 and 2. Essentially, we decided to discard the images where both effects were present to make it easier for the models to discern either of them. Moreover, some images contained a colorbar embedded within the image as shown in Figure 3 (unique ID: 13312). This was addressed by setting *image_artifacts* to 0 in order to exclude these images from the final dataset.

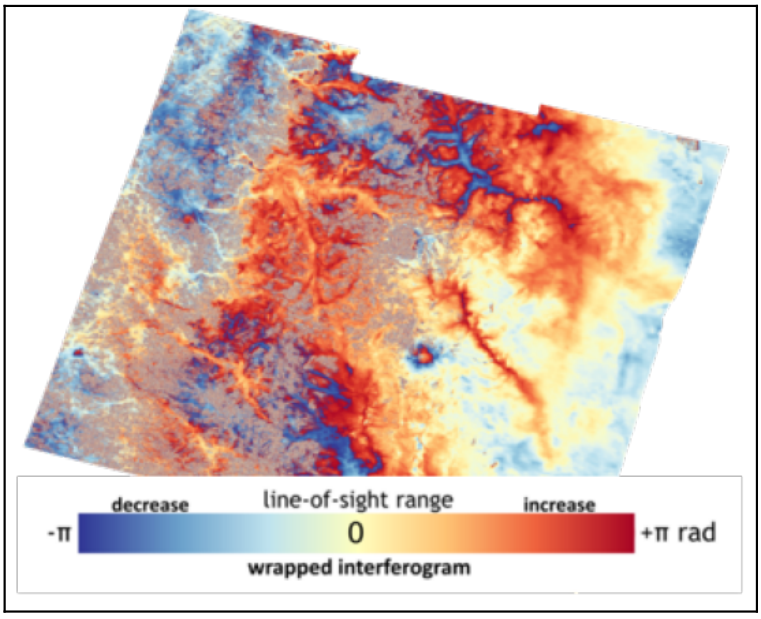

Figure 3. Vertical stratification caused by atmospheric delays can be detected in high altitude areas. No deformation activity can be detected. Image artifacts are detected on image.

These conditions were used to extract both the deforming and non-deforming images. The only additional requirement used to retrieve the deforming images was to set *is_crowd* to 0 as well as to

 doi:10.20944/preprints202605.0685.v1

discard images with 'Unidentified' for their *activity_type*. This excludes images that exhibit more than one fringe pattern related to deformation. It also makes sure that the deforming images included have their activity types known. All these restrictions put on the dataset essentially result in a cleaner and more usable dataset that has quality images and captions.

The Portable Network Graphics (PNG) images found in the Hephaestus dataset come in one of two modes. These are the Palette mode (P mode) and Red, Green, Blue, and Alpha mode (RGBA mode). P mode images come palettized which means there is essentially one channel for the image, and each pixel can take a value between 0 and 255 where each value designates a certain color from the palette. The general advantage of storing images this way is to save memory, as it requires a third of the space required of a typical RGB image. The disadvantage is being limited to only 256 unique colors, which could result in banding or artifacts. On the other hand, RGBA mode stores the image in 4 channels where the first 3 channels denote the different possible colors, and the fourth (Alpha) channel controls the color's transparency or opacity.

This is typically not an issue as one can convert from one mode to the other with ease. However, the main difference lies in the fact that the P mode images use an entirely different color scheme compared to the RGBA mode images. An example P mode image (unique ID: 16964) is shown in Figure 4, and an example RGBA image (unique ID: 8913) is shown in Figure 5. Although both image modes relay the same information just with a different colorbar, we decided to stick with one color scheme when building the dataset to be fed to the machine learning models. This was done to make it easier for the models to learn by not putting too much focus on the color scheme itself while also gaining the benefit of standardizing the image format.

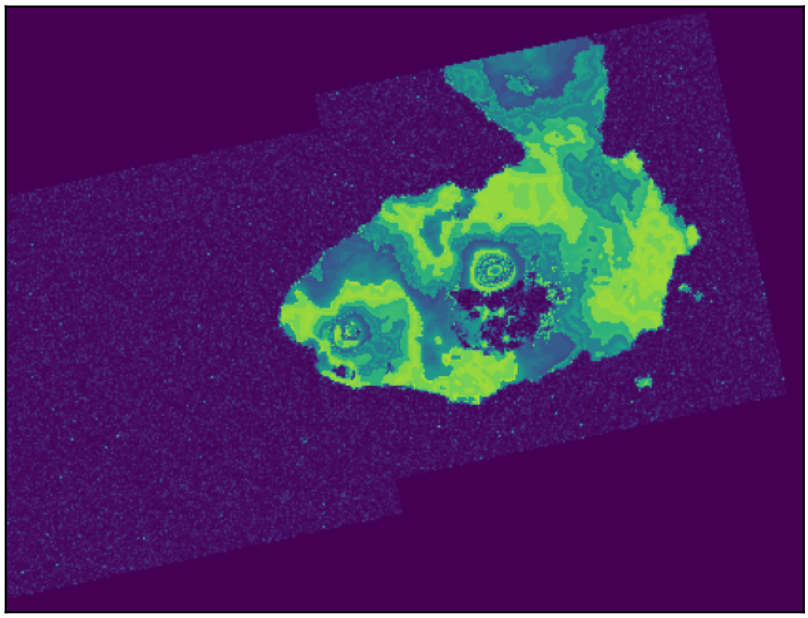

Figure 4. Vertical stratification effect can be detected on the left side of the region. A sill-type deformation pattern of medium intensity can be detected on the central side.

 doi:10.20944/preprints202605.0685.v1

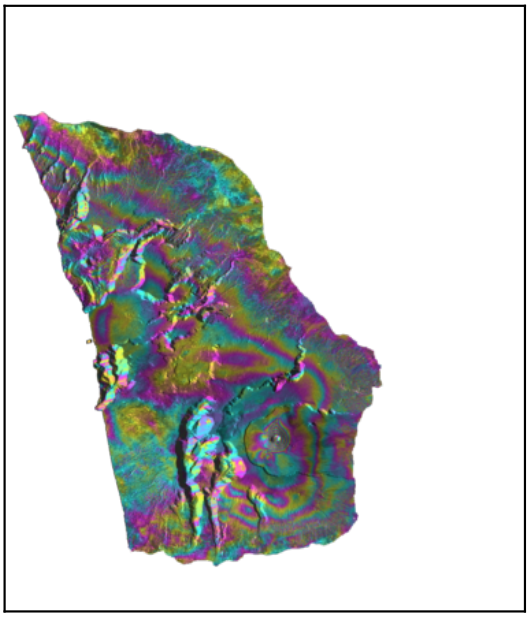

Figure 5. Turbulent mixing effect can be detected on the bottom-right, central and top-left side of the region. A sill-type deformation pattern of high intensity can be detected on the bottom-right side of the region.

After applying the conditions mentioned above to extract the clean images, we found that the majority of the resulting deforming images are in P mode. For example, the total deforming images came out to 634 after the preprocessing steps described, and 581 of those 634 images were in P mode, while 53 were in RGBA mode. The cleaned non-deforming images had roughly the same amount of P and RGBA modes (6139 P mode images and 5083 RGBA mode images). Due to this distribution, only the P images were extracted from the dataset and used.

As expected, there are far fewer deforming images compared to non-deforming images. In order to combat the class imbalance and ensure the models are capable of dealing with each class effectively, the number of non-deforming images used was truncated to more or less match the number of deforming images. This ensures an even amount of deforming and non-deforming images present in the training dataset. In fact, all of the 581 cleaned deforming images were used, and another 581 images from the cleaned non-deforming images were randomly selected to form a dataset consisting of 1162 images. For reproducibility purposes, this was done using Python's built-in random number generator, which was initialized with a fixed random seed of 17. Additionally, all extracted images were then converted from P mode to RGB mode in order to match the expected input format of the machine learning models to be used.

2.3. Machine Learning Models

2.3.1. Generative Image-to-Text Transformer Model

The Generative Image-to-Text Transformer (GIT) model is a vision-language model developed by Microsoft to perform image captioning and other image-conditioned language generation tasks such as question answering [67]. Unlike other models that depend on complex structures within their framework, GIT simplifies the architecture by using one image encoder and one text decoder as a single language modeling task.

At a high level, GIT operates by first extracting the visual features from the raw input image using the image encoder. This is followed by the text decoder that takes the flattened list of features and predicts the text description. The team leveraged large-scale image-text pairs t achieve the the goal of

 doi:10.20944/preprints202605.0685.v1

pre-training this vision language model. In fact, around 800 million image-text pairs were used for training, and the model achieved state-of-the-art performance on various image captioning tasks and question answering.

Architecturally, the image encoder was directly inspired by the previous work of Yuan et al. where a contrastive pre-trained model was built [68]. A feature projection module composed of a linear layer and a normalization layer is built on top of the encoder that projects the representations into the same embedding space as the language model tokens, ensuring compatibility with the transformer decoder. The text decoder itself follows a standard auto-regressive transformer architecture. It is made of several transformer blocks consisting of a self-attention layer and a feedforward layer. The attention mask used is such that the text token depends on the preceding tokens and all image tokens while allowing the image tokens to attend to each other. This unified architecture simplifies training and inference while maintaining strong performance across a variety of image-to-text tasks. Training begins with pre-training the image encoder by itself using the contrastive task. Then, using the generation task, the text decoder and image encoder are simultaneously pre-trained. The model architecture from the original paper is shown in Figure 6.

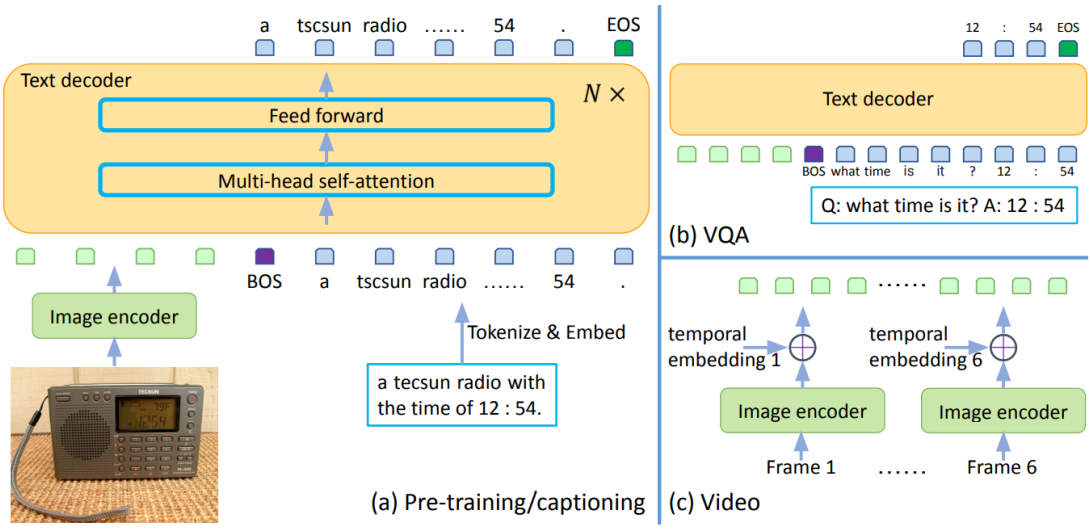

Figure 6. Network architecture of the GIT model for various language generation tasks.

In the context of scientific imagery such as InSAR data, GIT offers several advantages. To start with, the model's use of large-scale pre-training makes it suitable for situations where labeled data is limited, as domain-specific knowledge can be incorporated using the process of fine-tuning the model rather than training from scratch. Additionally, the generative nature allows for the production of flexible and form-free descriptions as opposed to a constrained set of candidate captions.

2.3.2. Bootstrapped Language-Image Pretraining Model

The Bootstrapping Language-Image Pretraining (BLIP) model is a vision-language framework that is capable of transferring flexibly to understanding and generation tasks [69]. The model is jointly pretrained with three complementary objectives in mind which are image-text contrastive learning (ITC), image-text matching (ITM), and image conditioned language modeling (LM). The contrastive objective aligns images and captions within a shared embedding space while encouraging semantically-related image-text pairs to have similar representations. The matching objective assists this alignment by distinguishing between paired and unpaired image-text combinations. Finally, the language modeling objective allows the model to generate natural language descriptions conditioned on the visual inputs. When combined, these objectives grant BLIP the ability to learn cross-modal correspondences and to generate captions fluently.

One of the key features of BLIP is its method of addressing noisy image-text pairs collected from the web. In fact, a bootstrapping method is employed where BLIP leverages its own generative

 doi:10.20944/preprints202605.0685.v1

capabilities to produce candidate captions using an image-grounded text decoder, which are then filtered and refined using an image-grounded text decoder to improve the quality of the training data. This iterative bootstrapping process enables a streamlined learning process with a lower noise ratio which directly contributes to BLIP's strong performance across benchmark datasets.

Its architectural design begins with a ViT as the image encoder that patches up and encodes an input image. Depending on the task configuration, BLIP uses a multimodal mixture of encoder-decoder. During image captioning training, the image-grounded text decoder is activated to generate textual descriptions in an autoregressive manner for the given image which is then optimized using cross-entropy loss. The full model architecture from the original paper is shown in Figure 7.

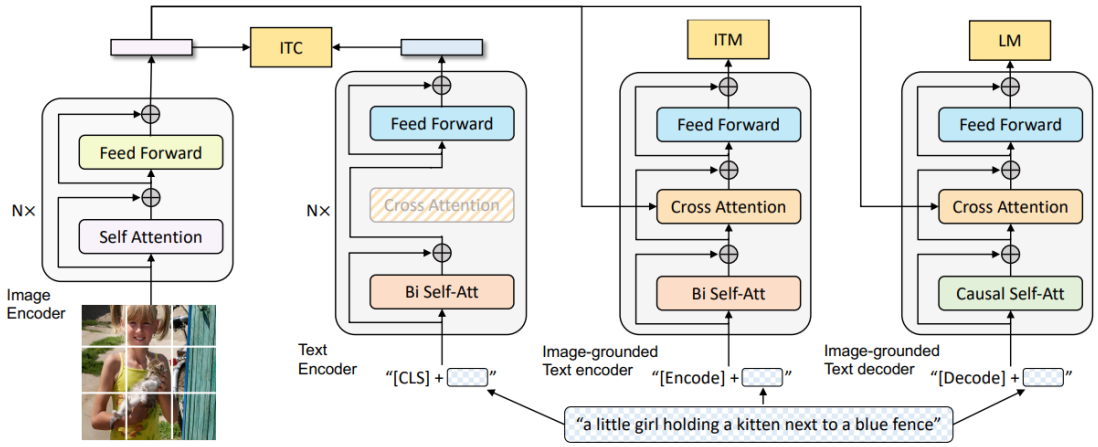

Figure 7. Network architecture of the BLIP model for various language generation tasks.

BLIP similarly offers advantages when it comes to domain-specific applications such as InSAR imagery. Its explicit modeling of image-text alignment and generation allows it to adapt to specialized visual semantics through the use of fine-tuning even in cases where the presence labeled data is limited. Moreover, the separation of visual encoding and language decoding provides greater control over cross-modal interactions which can be beneficial when transferring models trained on natural images to non-natural imagery. However, as with other pretrained vision-language models, the performance of BLIP in such domains will depend on the extent to which fine-tuning can overcome biases introduced during the large-scale pre-training.

2.3.3. Retrieval-Based Model

In addition to the generative captioning approaches described above, we implement a retrieval-based image-text baseline model that maps InSAR images into a visual feature space and predicts captions by nearest-neighbor matching. Rather than generating captions autoregressively, this method frames captioning as example retrieval: a query image is compared against a bank of training images, and the caption attached to the most similar training image is returned as the prediction. This setup provides a direct way to assess whether pre-trained visual representations capture semantics relevant to our domain.

The model uses the vision encoder from a pre-trained BLIP captioning model (Salesforce/blip-image-captioning-base) together with its corresponding processor. For each image, we extract the pooled output of the BLIP vision backbone and use that vector as the image embedding. In contrast to the contrastive variant, this retrieval pipeline does not include a separate text encoder, projection heads, or any cross-modal alignment loss. Captions are used only as labels associated with training images.

The training split is therefore used to build an embedding index rather than to optimize parameters. We compute and store embeddings for all training images in advance, while keeping the pre-trained BLIP weights fixed. This design keeps the method simple and reduces the risk of overfitting on a limited InSAR dataset, since no additional learned components are introduced.

 doi:10.20944/preprints202605.0685.v1

During inference, each test image is passed through the same BLIP vision encoder to obtain a query embedding. We then compute cosine similarity between that query vector and all stored training-image embeddings. The index of the highest-similarity training image is selected, and its caption is transferred as the predicted caption for the test sample.

Evaluation is performed by comparing the retrieved caption with the ground-truth caption for each test image and reporting the retrieval accuracy over the test split. Because predictions are copied from existing training captions rather than generated token by token, outputs are deterministic and avoid the hallucination behavior often observed in generative models. This can be advantageous for scientific imagery where faithful and conservative descriptions are preferred.

Overall, this nearest-neighbor retrieval model serves as a lightweight and interpretable baseline alongside generative captioning methods. It isolates how far visual similarity alone can support caption prediction in the InSAR setting, and it provides a transparent reference point for judging the added value of more complex generation-based approaches.

3. Results

The baseline retrieval model does not require training, as predictions are generated by selecting the caption associated with the most similar image in the embedding space using cosine similarity. Due to the nature of this model, we decided to evaluate performance based on its raw accuracy, where the raw accuracy is defined as whether or not there is an exact match (string-to-string) of the true caption and the retrieved caption. It is a rather unforgiving metric, but it fits the scope of this method, as it is very quick to implement and deploy.

To build the embedding index, 90% of the dataset was used, and 10% was left for inference purposes. This was done using Hugging Face's API [70] and a fixed random seed of 17. The model achieved a raw accuracy of 43.59% on the test set consisting of 117 images. Representative correct and incorrect predictions are shown in Figures 8, 9, 10 and 11, respectively.

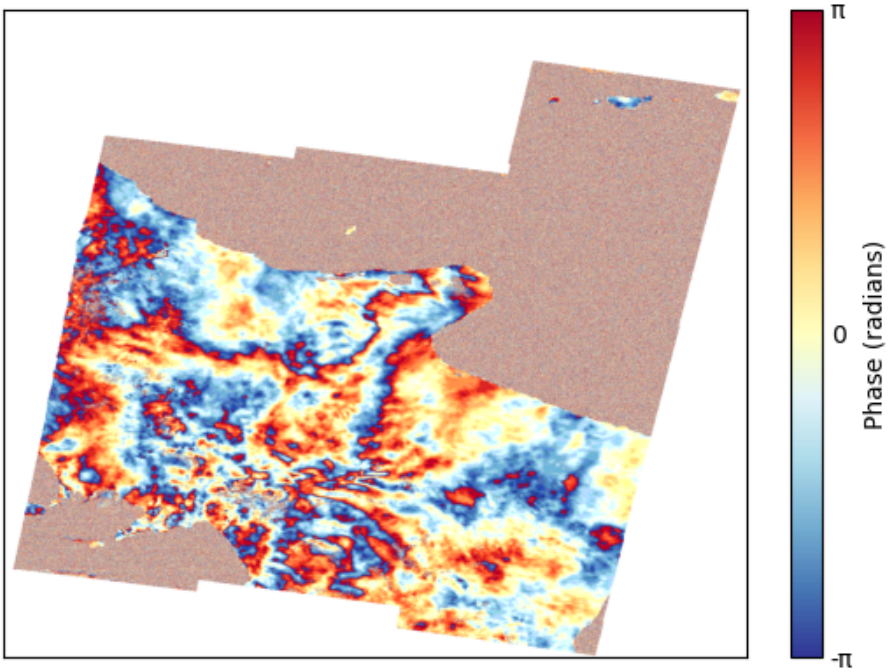

Figure 8. Matched caption: Turbulent mixing effect or wave-like patterns caused by liquid and solid particles of the atmosphere can be detected around the area. No deformation activity can be detected.

 doi:10.20944/preprints202605.0685.v1

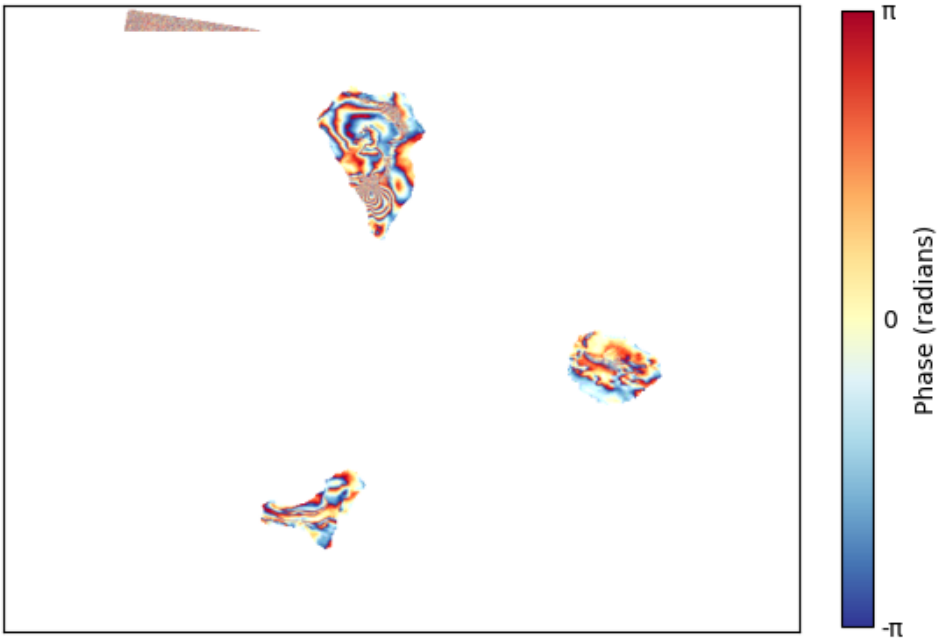

Figure 9. Matched caption: Turbulent mixing effect can be detected on the top, right and bottom side. A dyke-type deformation pattern of high intensity can be detected on the top side of the image.

 doi:10.20944/preprints202605.0685.v1

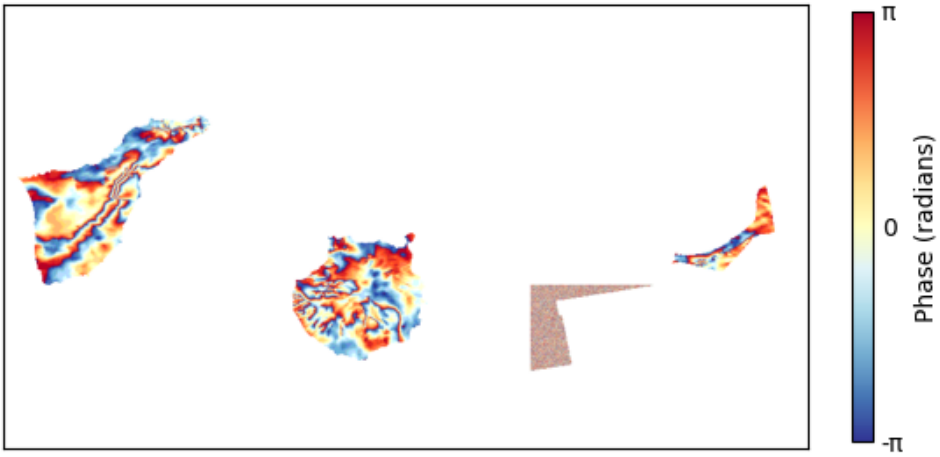

Figure 10. True caption: Vertical stratification caused by atmospheric delays can be detected in high altitude areas. No deformation activity can be detected. Retrieved caption: Turbulent mixing effect or wave-like patterns caused by liquid and solid particles of the atmosphere can be detected around the area. No deformation activity can be detected.

 doi:10.20944/preprints202605.0685.v1

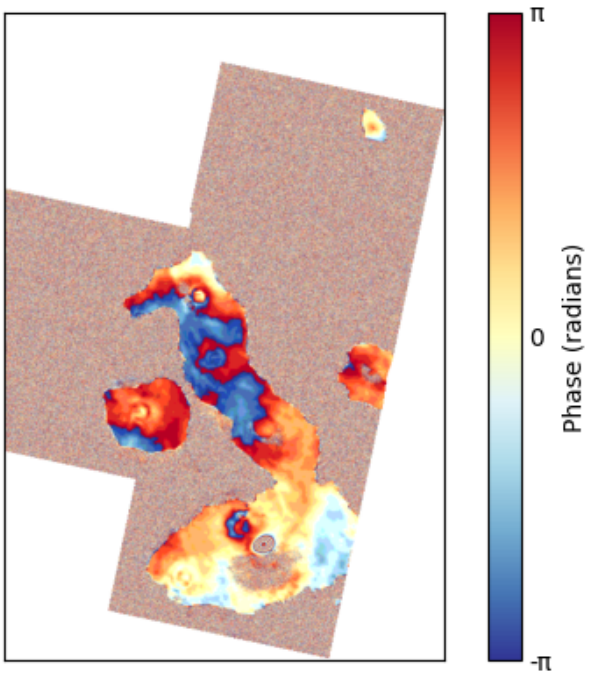

Figure 11. True caption: A sill-type deformation pattern of medium intensity can be detected at the bottom of the image. The atmospheric impact is very low. Retrieved caption: Turbulent mixing effect can be detected on the leftmost and bottom-left side of the region. A sill-type deformation pattern of high intensity can be detected at the bottom of the image.

For the generative models, a similar data split was used, with an additional 10% of the training portion set aside for validation purposes. Relying on raw accuracy as the metric for the generative approach is too unforgiving as slight semantic differences between generated and ground truth captions wouldn't be counted positively. Instead, we chose to rely on a different set of metrics which are the Rouge-1, Rouge-2, Rouge-L, and BLEU scores.

Training of the models was done using UC Davis's high-performance computing (HPC) cluster called Hive. Model performance was monitored on the validation set across epochs. Plots showing the performances of the GIT and BLIP models are shown in Figures 12 and 13 respectively. Both models were trained for 30 epochs with evaluation metrics stabilizing after approximately 25 epochs.

 doi:10.20944/preprints202605.0685.v1

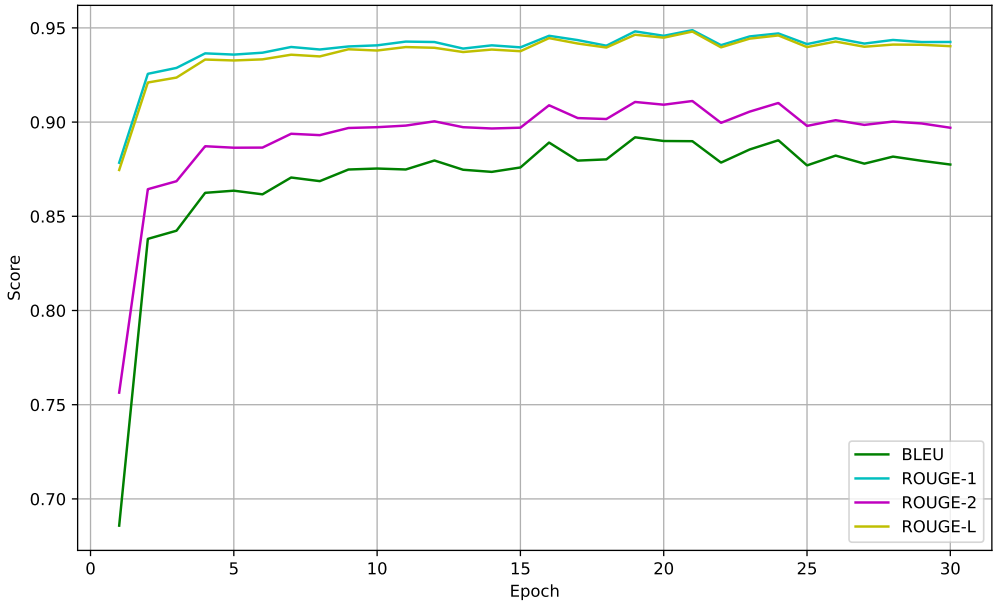

Figure 12. Metric scores tracked for the GIT model over the epochs during training.

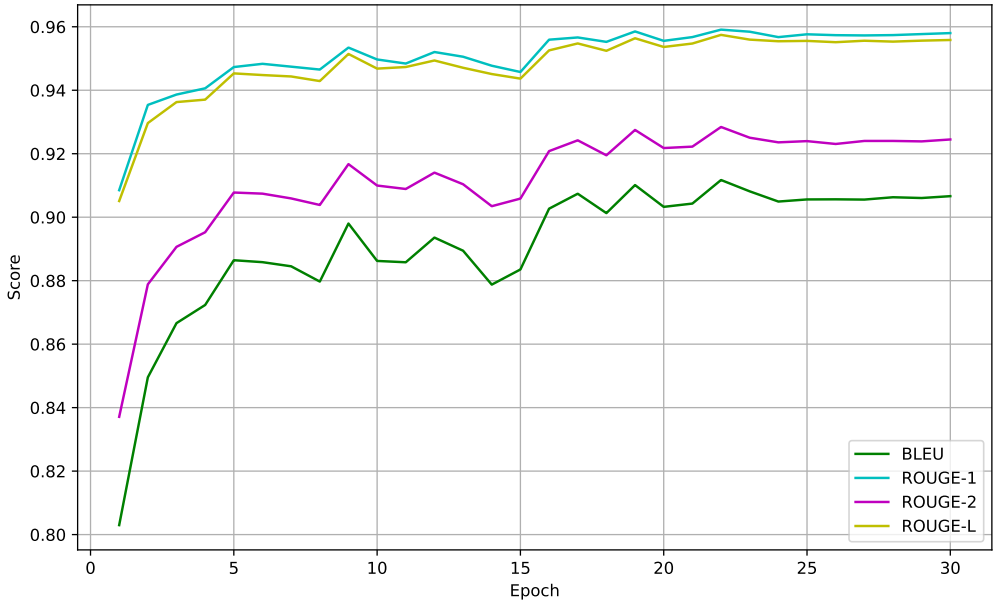

Figure 13. Metric scores tracked for the BLIP model over the epochs during training.

The models are set into evaluation mode in order to perform inference on the test set. A few decoding parameters were tuned to optimize the caption generation process. A beam search with a width of 5 was used instead of a greedy approach. A temperature of 0.7 was applied to control the sharpness of the token distribution. Moreover, top_k sampling ($k = 50$) and nucleus sampling ($p = 0.9$) were employed to balance diversity and coherence in the generated captions. Example outputs along with their images are presented in Figures 14–17.

 doi:10.20944/preprints202605.0685.v1

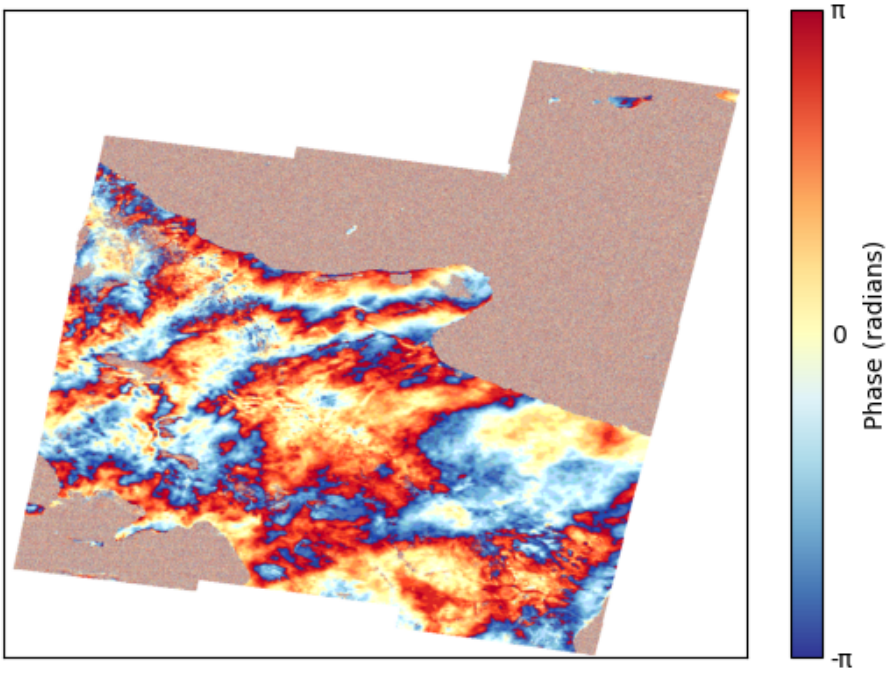

Figure 14. True caption: Turbulent mixing effect or wave-like patterns caused by liquid and solid particles of the atmosphere can be detected around the area. No deformation activity can be detected. BLIP caption: turbulent mixing effect or wave - like patterns caused by liquid and solid particles of the atmosphere can be detected around the area. no deformation activity. GIT caption: turbulent mixing effect can be detected on the bottom - right and central - left side of the region. a sill - type deformation pattern of high intensity is detected at the bottom of the image. there is noise in deformation zone due to high temporal baseline between the two images.

 doi:10.20944/preprints202605.0685.v1

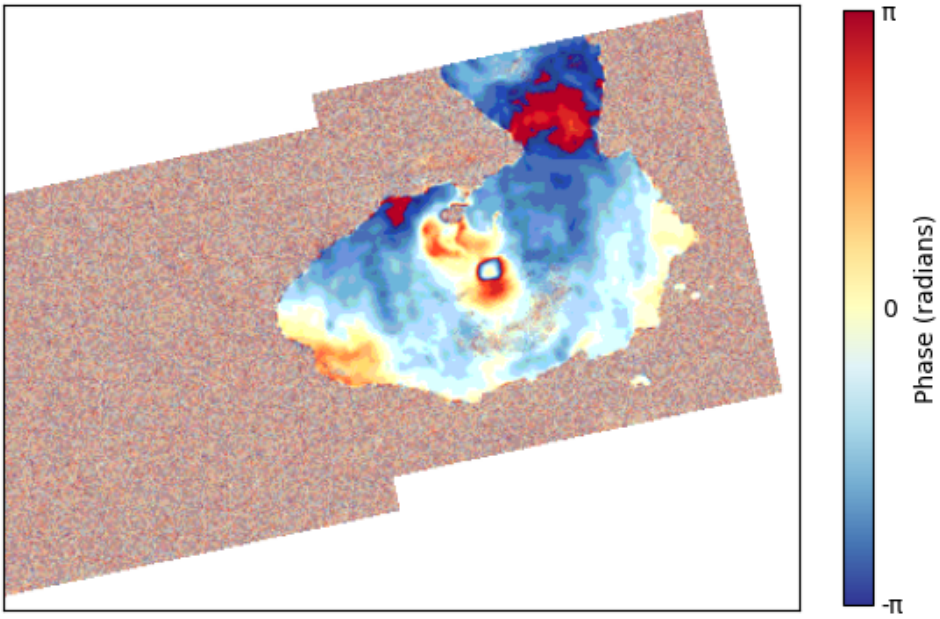

Figure 15. True caption: A sill-type deformation pattern of low intensity can be detected on the central side. The atmospheric impact is low. BLIP caption: a sill - type deformation pattern of low intensity can be detected on the central side. the atmospheric impact is low. GIT caption: vertical stratification effect is detected on the high altitude areas of the region. a sill - type deformation pattern of low intensity is detected at the bottom of the image.

 doi:10.20944/preprints202605.0685.v1

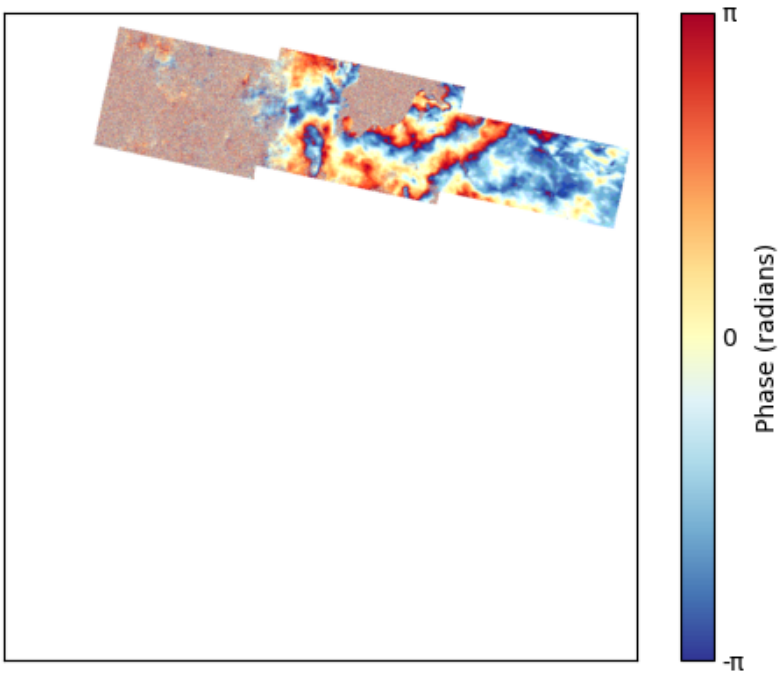

Figure 16. True caption: Turbulent mixing effect or wave-like patterns caused by liquid and solid particles of the atmosphere can be detected around the area. No deformation activity can be detected. BLIP caption: turbulent mixing effect or wave - like patterns caused by liquid and solid particles of the atmosphere can be detected around the area. no deformation activity. GIT caption: vertical stratification effect is detected on the high altitude areas of the region. a sill - type deformation pattern of medium intensity is detected at the bottom of the image.

 doi:10.20944/preprints202605.0685.v1

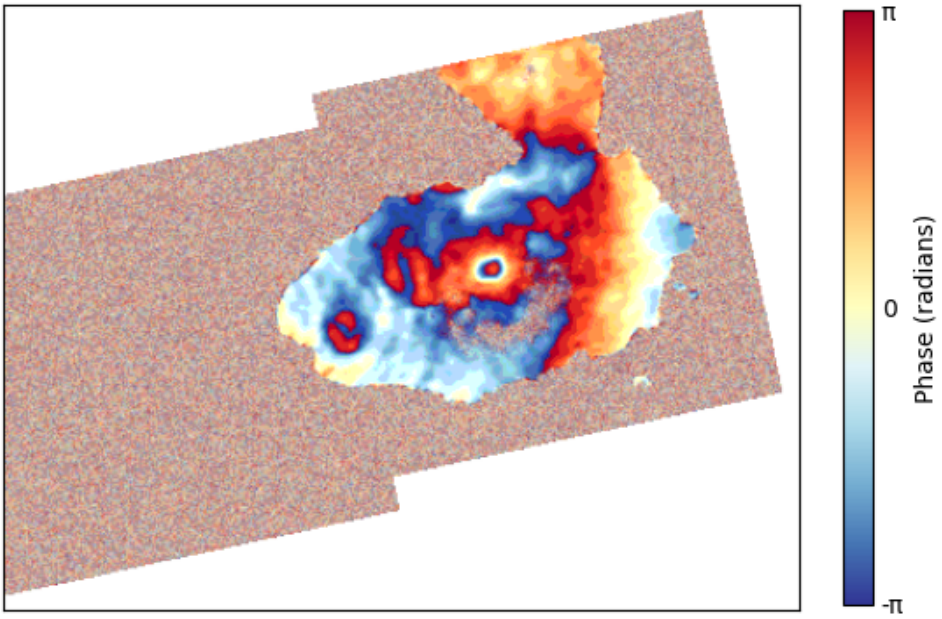

Figure 17. True caption: A sill-type deformation pattern of low intensity can be detected on the central side. The atmospheric impact is low. BLIP caption: a sill - type deformation pattern of low intensity can be detected on the central side. the atmospheric impact is low. GIT caption: vertical stratification effect is detected on the high altitude areas of the region. a sill - type deformation pattern of low intensity is detected at the bottom of the image.

4. Discussion

4.1. Model Performance

Among the generative models, BLIP consistently achieves slightly superior performance as reflected by higher evaluation metric scores. Additionally, comparison with the retrieval-based model reveals that there are some examples in the test that the retrieval model gets right where the generative model fails and vice versa.

At first glance, the retrieval model appears competitive in a superior way, especially given the minimal computational requirements. However, this performance is partly attributable to its inability to produce novel captions but instead retrieves existing captions from visually similar images. Consequently, its effectiveness is strongly influenced by the caption distribution within the dataset.

In particular, the caption uniqueness distribution of the underlying dataset being used introduces a notable bias. The plot showing this can be seen in Figure 18. Despite splitting up the dataset such that there is roughly an equal amount of deforming and non-deforming images, the uniqueness of each plays a critical role in the model embedding, and the diversity of captions differs substantially between these classes. To elaborate, when it comes to the non-deforming images, there are practically little to no unique captions. In fact, the most common non-deformation caption is: "Turbulent mixing effect or wave-like patterns caused by liquid and solid particles of the atmosphere can be detected around the area. No deformation activity can be detected." with a count of 8141. The most common deforming caption which is: "A sill-type deformation pattern of low intensity can be detected on the central side. The atmospheric impact is low." has a count of just 25, so there are many more unique deforming captions than there are unique non-deforming captions. This imbalance favors retrieval-based methods, as repeated captions increase the likelihood of exact matches.

 doi:10.20944/preprints202605.0685.v1

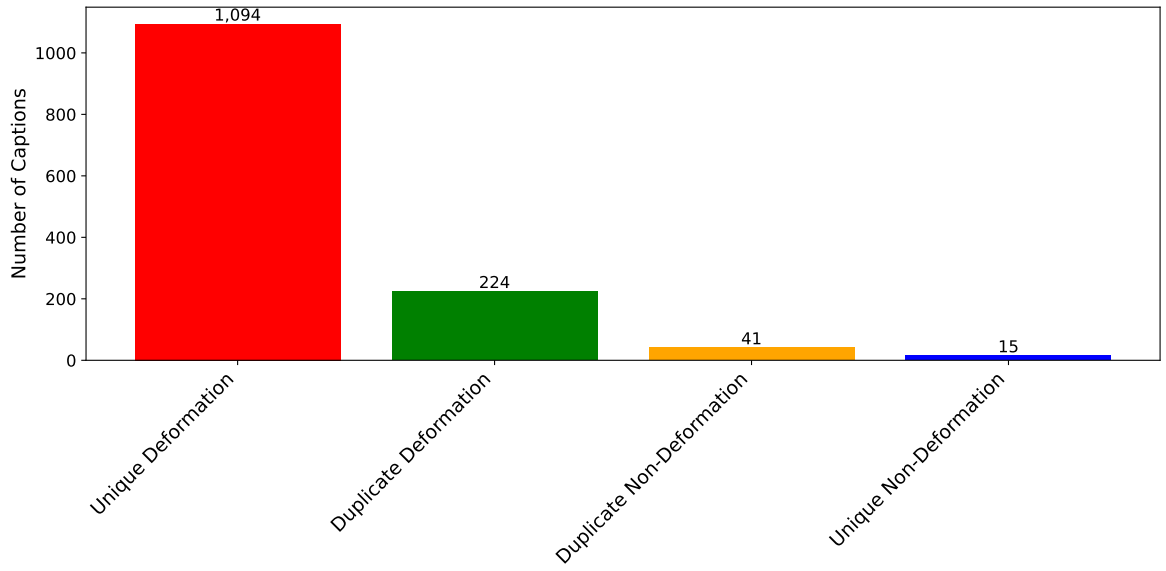

Figure 18. Bar plot showing the distribution of captions in the dataset by type and uniqueness.

Given these observations, a hybrid approach that combines generative and retrieval-based methods is likely to be most effective. In addition, incorporating human oversight for validation further improves reliability. Such a system could substantially accelerate the annotation of large-scale InSAR datasets while maintaining scientific accuracy.

4.2. Limitations

Despite the promising results obtained in this study, a few limitations should be acknowledged. First, the size of the dataset used for training and evaluation is relatively small compared to the large-scale datasets that are typically used for training modern vision-language models. Although the models used in this work benefit from the extensive pre-training, the limited number of labeled InSAR images may restrict their ability to fully capture the diversity of deformation patterns and atmospheric artifacts that are present in these real-world observations.

Second, the pre-trained models used in this study were trained on large collections of natural images paired with general language descriptions. As a result, these models may retain biases and semantic priors that are not necessarily fully aligned with the characteristics of scientific imagery such as InSAR data. Even though fine-tuning helps adapt the models to this domain, the mismatch between natural image distributions and specialized geophysical imagery may still influence the quality and accuracy of the resulting captions.

Third, evaluating the quality of the generated captions remains a challenging problem. Automated metrics commonly used in image captioning, such as BLEU or related measures, do not necessarily capture whether a caption accurately reflects the underlying geophysical interpretation of the image. Arguably, the best way to measure the quality of the caption for an image is still carefully designing a human evaluation campaign in which multiple users score the produced sentences [1,71]. In practice, this means that a powerful evaluation would require human assessment by domain experts who can judge whether the descriptions correctly identify deformation patterns and atmospheric effects.

Finally, the retrieval-based approach explored in this work is inherently limited by its reliance on existing captions within the dataset. While retrieval methods can produce grammatically consistent and scientifically plausible descriptions, they cannot generate novel captions for previously unseen patterns. This restricts the flexibility of the approach when encountering deformation signals that differ from those present in the training set.

 doi:10.20944/preprints202605.0685.v1

4.3. Future Directions

Several directions could be taken to further improve the performance and applicability of the automated InSAR image captioning system. First, one could expand the the effective dataset size through domain-specific data augmentation strategies. To illustrate, deformation images could be geometrically transformed through rotations or flips to generate additional training samples. However, applying such transformations requires careful modification of the corresponding captions in order to keep the spatial descriptions consistent with the transformed image. Developing automated procedures for adjusting captions after geometric transformations could significantly increase the diversity of training data while preserving semantic correctness.

Another direction involves exploring multi-label prediction frameworks that directly infer the structured annotations associated with each interferogram. Since the Hephaestus dataset contains multiple labels describing atmospheric effects, deformation type, intensity, and other attributes, a model trained to predict these labels would provide a more interpretable intermediate representation of the image. Combining these predicted labels to construct the natural language caption would potentially improve both accuracy and consistency of the resulting descriptions.

Another potential improvement involves incorporating specialized feature extraction methods that are specifically tailored to geophysical imagery. For example, including deformation boundary coordinates or segmentation masks as additional inputs could help guide the models' attention towards the physically meaningful regions of the InSAR image. This information may allow the model to better capture the spatial structure of deformation patterns and reduce the influence of atmospheric artifacts.

Finally, future work may also benefit from using the higher fidelity image representations that retain the underlying phase information of the interferograms. Rather than relying on the compressed PNG images, models could be trained directly on the TIFF files containing the raw phase data. Although this approach would increase memory and computational requirements, it may allow the models to learn more precise representations of deformation patterns and improve captioning performance.

5. Conclusions

In this paper, we investigated the feasibility of applying modern vision–language models to the task of automatically generating textual descriptions for InSAR imagery. After preprocessing the Hephaestus dataset and curating a training dataset, we evaluated both generative captioning models and a retrieval-based approach designed to align image and text representations within a shared embedding space. The models, pre-trained on natural images, were fine-tuned on domain-specific data in order to learn the relationships between visual patterns in interferograms and their corresponding descriptive captions.

The results demonstrate that pretrained vision–language models can be adapted to specialized scientific imagery despite being originally trained on natural image datasets. Both generative and retrieval-based approaches were capable of identifying meaningful features within the interferograms and producing captions that capture important aspects of deformation and atmospheric effects. While several limitations remain, including dataset size and evaluation challenges, the findings suggest that automated captioning systems may provide a useful tool for assisting researchers in the interpretation of large volumes of satellite radar imagery. Continued advances in domain-specific training data, model architectures, and evaluation strategies may further improve the reliability and usefulness of such systems for geophysical monitoring applications.

Author Contributions: Conceptualization, J.Y.; methodology, J.Y.; software, J.Y.; validation, J.Y.; formal analysis, J.Y.; investigation, J.Y.; resources, J.B.R.; data curation, J.Y.; writing—original draft preparation, J.Y.; writing—review and editing, J.Y.; visualization, J.Y.; supervision, J.B.R.; project administration, J.B.R.; funding acquisition, J.B.R. All authors have read and agreed to the published version of the manuscript.

 doi:10.20944/preprints202605.0685.v1

Funding: This research was funded under the United States Department of Energy grant DE-SC0017324 to the University of California, Davis.

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Data Availability Statement: The authors of the Hephaestus dataset have published it and maintain it on: https://github.com/Orion-AI-Lab/Hephaestus

Acknowledgments: LiCSAR contains modified Copernicus Sentinel data [2014–2021] analysed by the Centre for the Observation and Modelling of Earthquakes, Volcanoes and Tectonics (COMET). LiCSAR uses JASMIN, the UK's collaborative data analysis environment (http://jasmin.ac.uk) (accessed on 15 March 2025). The authors have reviewed and edited the output and take full responsibility for the content of this publication. The authors acknowledge the High Performance Computing Core Facility at the University of California, Davis, for providing computational resources that have contributed to the research results reported in this paper.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

CNN	Convolutional Neural Network
RNN	Recurrent Neural Network
LSTM	Long Short Term Memory
ViT	Vision Transformer
UAVs	Unmanned Aerial Vehicles
InSAR	Interferometric Synthetic Aperture Radar
PNG	Portable Network Graphics
P	Palette
RGBA	Red, Green, Blue, and Alpha
GIT	Generative Image-to-Text Transformer
InfoNCE	Information Noise-Contrastive Estimation
ITC	Image-Text Contrastive Learning
ITM	Image-Text Matching
LM	Image Conditioned Language Modeling

References

1. Stefanini, M.; Cornia, M.; Baraldi, L.; Cascianelli, S.; Fiameni, G.; Cucchiara, R. From show to tell: A survey on deep learning-based image captioning. *IEEE transactions on pattern analysis and machine intelligence* **2022**, *45*, 539–559.
2. He, X.; Deng, L. Deep learning for image-to-text generation: A technical overview. *IEEE Signal Processing Magazine* **2017**, *34*, 109–116.
3. Kumar, A.; Goel, S. A survey of evolution of image captioning techniques. *International Journal of Hybrid Intelligent Systems* **2017**, *14*, 123–139.
4. Li, L.J.; Fei-Fei, L. What, where and who? classifying events by scene and object recognition. In Proceedings of the 2007 IEEE 11th international conference on computer vision. IEEE, 2007, pp. 1–8.
5. Li, L.J.; Socher, R.; Fei-Fei, L. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2009, pp. 2036–2043.
6. Felzenszwalb, P.F.; Girshick, R.B.; McAllester, D.; Ramanan, D. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence* **2009**, *32*, 1627–1645.
7. Yang, Y.; Teo, C.; Daumé III, H.; Aloimonos, Y. Corpus-guided sentence generation of natural images. In Proceedings of the Proceedings of the 2011 conference on empirical methods in natural language processing, 2011, pp. 444–454.

 doi:10.20944/preprints202605.0685.v1

8. Farhadi, A.; Hejrati, M.; Sadeghi, M.A.; Young, P.; Rashtchian, C.; Hockenmaier, J.; Forsyth, D. Every picture tells a story: Generating sentences from images. In Proceedings of the European conference on computer vision. Springer, 2010, pp. 15–29.
9. Li, S.; Kulkarni, G.; Berg, T.; Berg, A.; Choi, Y. Composing simple image descriptions using web-scale n-grams. In Proceedings of the Proceedings of the fifteenth conference on computational natural language learning, 2011, pp. 220–228.
10. Kulkarni, G.; Premraj, V.; Ordonez, V.; Dhar, S.; Li, S.; Choi, Y.; Berg, A.C.; Berg, T.L. Babytalk: Understanding and generating simple image descriptions. *IEEE transactions on pattern analysis and machine intelligence* **2013**, *35*, 2891–2903.
11. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W.J. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318.
12. Lin, C.Y.; Hovy, E. Automatic evaluation of summaries using n-gram co-occurrence statistics. In Proceedings of the Proceedings of the 2003 human language technology conference of the North American chapter of the association for computational linguistics, 2003, pp. 150–157.
13. Mitchell, M.; Dodge, J.; Goyal, A.; Yamaguchi, K.; Stratos, K.; Han, X.; Mensch, A.; Berg, A.; Berg, T.; Daumé III, H. Midge: Generating image descriptions from computer vision detections. In Proceedings of the Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics, 2012, pp. 747–756.
14. Hodosh, M.; Young, P.; Hockenmaier, J. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research* **2013**, *47*, 853–899.
15. Kiros, R.; Salakhutdinov, R.; Zemel, R.S. Unifying visual-semantic embeddings with multimodal neural language models. *arXiv preprint arXiv:1411.2539* **2014**.
16. Lu, Z.; Li, H. A deep architecture for matching short texts. *Advances in neural information processing systems* **2013**, *26*.
17. Gong, Y.; Jia, Y.; Leung, T.; Toshev, A.; Ioffe, S. Deep convolutional ranking for multilabel image annotation. *arXiv preprint arXiv:1312.4894* **2013**.
18. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* **2012**, *25*.
19. Socher, R.; Karpathy, A.; Le, Q.V.; Manning, C.D.; Ng, A.Y. Grounded compositional semantics for finding and describing images with sentences. *Transactions of the Association for Computational Linguistics* **2014**, *2*, 207–218.
20. Chen, X.; Lawrence Zitnick, C. Mind's eye: A recurrent visual representation for image caption generation. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 2422–2431.
21. Banerjee, S.; Lavie, A. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In Proceedings of the Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization, 2005, pp. 65–72.
22. Vedantam, R.; Lawrence Zitnick, C.; Parikh, D. Cider: Consensus-based image description evaluation. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 4566–4575.
23. Vinyals, O.; Toshev, A.; Bengio, S.; Erhan, D. Show and tell: A neural image caption generator. In Proceedings of the Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3156–3164.
24. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Advances in neural information processing systems* **2017**, *30*.
25. Yang, X.; Zhang, H.; Cai, J. Learning to collocate neural modules for image captioning. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 4250–4260.
26. Li, G.; Zhu, L.; Liu, P.; Yang, Y. Entangled transformer for image captioning. In Proceedings of the Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 8928–8937.
27. Pan, Y.; Yao, T.; Li, Y.; Mei, T. X-linear attention networks for image captioning. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 10971–10980.
28. Cornia, M.; Stefanini, M.; Baraldi, L.; Cucchiara, R. Meshed-memory transformer for image captioning. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2020, pp. 10578–10587.

 doi:10.20944/preprints202605.0685.v1

29. Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* **2020**.
30. Liu, W.; Chen, S.; Guo, L.; Zhu, X.; Liu, J. Cptr: Full transformer network for image captioning. *arXiv preprint arXiv:2101.10804* **2021**.
31. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the International conference on machine learning. PmLR, 2021, pp. 8748–8763.
32. Wang, Z.; Yu, J.; Yu, A.W.; Dai, Z.; Tsvetkov, Y.; Cao, Y. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904* **2021**.
33. Staniūtė, R.; Šešok, D. A systematic literature review on image captioning. *Applied Sciences* **2019**, *9*, 2024.
34. Mao, J.; Xu, W.; Yang, Y.; Wang, J.; Huang, Z.; Yuille, A. Deep captioning with multimodal recurrent neural networks (m-rnn). *arXiv preprint arXiv:1412.6632* **2014**.
35. Xiao, B.; Wang, Y.; Kang, S.C. Deep learning image captioning in construction management: a feasibility study. *Journal of Construction Engineering and Management* **2022**, *148*, 04022049.
36. Lee, H.; Cho, H.; Park, J.; Chae, J.; Kim, J. Cross encoder-decoder transformer with global-local visual extractor for medical image captioning. *Sensors* **2022**, *22*, 1429.
37. Gamidi, R.; Hemasri, M.; Muppala, T.; Chowdary, V.; Palaniswamy, S.; et al. Enhancing Underwater Image Captioning Using Transformer Models and Augmented Terrestrial Datasets. In Proceedings of the 2025 International Conference on Pervasive Computational Technologies (ICPCT). IEEE, 2025, pp. 944–949.
38. Zhang, K.; Li, P.; Wang, J. A review of deep learning-based remote sensing image caption: Methods, models, comparisons and future directions. *Remote Sensing* **2024**, *16*, 4113.
39. Zhao, B. A systematic survey of remote sensing image captioning. *IEEE Access* **2021**, *9*, 154086–154111.
40. Sharma, H.; Padha, D. Domain-specific image captioning: a comprehensive review. *International Journal of Multimedia Information Retrieval* **2024**, *13*, 20.
41. Qu, B.; Li, X.; Tao, D.; Lu, X. Deep semantic understanding of high resolution remote sensing image. In Proceedings of the 2016 International conference on computer, information and telecommunication systems (Cits). IEEE, 2016, pp. 1–5.
42. Cigna, F.; Tapete, D.; Lu, Z. Remote sensing of volcanic processes and risk, 2020.
43. Tronin, A.A. Remote sensing and earthquakes: A review. *Physics and Chemistry of the Earth, parts A/B/C* **2006**, *31*, 138–142.
44. Tralli, D.M.; Blom, R.G.; Zlotnicki, V.; Donnellan, A.; Evans, D.L. Satellite remote sensing of earthquake, volcano, flood, landslide and coastal inundation hazards. *ISPRS Journal of Photogrammetry and Remote Sensing* **2005**, *59*, 185–198.
45. Osmanoğlu, B.; Sunar, F.; Wdowinski, S.; Cabral-Cano, E. Time series analysis of InSAR data: Methods and trends. *Isprs journal of photogrammetry and remote sensing* **2016**, *115*, 90–102.
46. Biggs, J.; Bergman, E.; Emmerson, B.; Funning, G.J.; Jackson, J.; Parsons, B.; Wright, T.J. Fault identification for buried strike-slip earthquakes using InSAR: The 1994 and 2004 Al Hoceima, Morocco earthquakes. *Geophysical Journal International* **2006**, *166*, 1347–1362.
47. Yazbeck, J.; Rundle, J.B. A Fusion of Geothermal and InSAR Data with Machine Learning for Enhanced Deformation Forecasting at the Geysers. *Land* **2023**, *12*, 1977.
48. Yazbeck, J.; Rundle, J.B. Predicting short-term deformation in the central valley using machine learning. *Remote Sensing* **2023**, *15*, 449.
49. Anantrasirichai, N.; Biggs, J.; Albino, F.; Bull, D. A deep learning approach to detecting volcano deformation from satellite imagery using synthetic datasets. *Remote Sensing of Environment* **2019**, *230*, 111179.
50. Bountos, N.I.; Michail, D.; Papoutsis, I. Learning from synthetic InSAR with vision transformers: The case of volcanic unrest detection. *IEEE Transactions on Geoscience and Remote Sensing* **2022**, *60*, 1–12.
51. Wu, Y.Y.; Madson, A. Error sources of interferometric synthetic aperture radar satellites. *Remote Sensing* **2024**, *16*, 354.
52. Rosen, P.A.; Hensley, S.; Joughin, I.R.; Li, F.K.; Madsen, S.N.; Rodriguez, E.; Goldstein, R.M. Synthetic aperture radar interferometry. *Proceedings of the IEEE* **2002**, *88*, 333–382.
53. Ferretti, A.; Monti-Guarnieri, A.; Prati, C.; Rocca, F.; Massonet, D. *InSAR principles-guidelines for SAR interferometry processing and interpretation*; Vol. 19, 2007.
54. Bürgmann, R.; Rosen, P.A.; Fielding, E.J. Synthetic aperture radar interferometry to measure Earth's surface topography and its deformation. *Annual review of earth and planetary sciences* **2000**, *28*, 169–209.

 doi:10.20944/preprints202605.0685.v1

55. Bountos, N.I.; Papoutsis, I.; Michail, D.; Karavias, A.; Elias, P.; Parcharidis, I. Hephaestus: A large scale multitask dataset towards InSAR understanding. In Proceedings of the Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 1453–1462.
56. Lazeckỳ, M.; Spaans, K.; González, P.J.; Maghsoudi, Y.; Morishita, Y.; Albino, F.; Elliott, J.; Greenall, N.; Hatton, E.; Hooper, A.; et al. LiCSAR: An automatic InSAR tool for measuring and monitoring tectonic and volcanic activity. *Remote Sensing* **2020**, *12*, 2430.
57. Morishita, Y.; Lazecky, M.; Wright, T.J.; Weiss, J.R.; Elliott, J.R.; Hooper, A. LiCSBAS: An open-source InSAR time series analysis package integrated with the LiCSAR automated Sentinel-1 InSAR processor. *Remote Sensing* **2020**, *12*, 424.
58. Lawrence, B.N.; Bennett, V.L.; Churchill, J.; Juckes, M.; Kershaw, P.; Pascoe, S.; Pepler, S.; Pritchard, M.; Stephens, A. Storing and manipulating environmental big data with JASMIN. In Proceedings of the 2013 IEEE international conference on big data. IEEE, 2013, pp. 68–75.
59. Kiyoo, M. Relations between the eruptions of various volcanoes and the deformations of the ground surfaces around them. *Earthq Res Inst* **1958**, *36*, e134.
60. Milczarek, W.; Kopeć, A.; Głąbicki, D.; Bugajska, N. Induced seismic events—distribution of ground surface displacements based on InSAR methods and Mogi and Yang models. *Remote Sensing* **2021**, *13*, 1451.
61. Gudmundsson, A. How local stresses control magma-chamber ruptures, dyke injections, and eruptions in composite volcanoes. *Earth-science reviews* **2006**, *79*, 1–31.
62. Okada, Y. Surface deformation due to shear and tensile faults in a half-space. *Bull. Seismol. Soc. Am* **1985**, *75*, 1135–1154.
63. Fialko, Y.; Khazan, Y.; Simons, M. Deformation due to a pressurized horizontal circular crack in an elastic half-space, with applications to volcano geodesy. *Geophysical Journal International* **2001**, *146*, 181–190.
64. Giudicepietro, F.; Macedonio, G.; Martini, M. A physical model of sill expansion to explain the dynamics of unrest at calderas with application to Campi Flegrei. *Frontiers in Earth Science* **2017**, *5*, 54.
65. Yang, X.M.; Davis, P.M.; Dieterich, J.H. Deformation from inflation of a dipping finite prolate spheroid in an elastic half-space as a model for volcanic stressing. *Journal of Geophysical Research: Solid Earth* **1988**, *93*, 4249–4257.
66. Galbusera, F.; Cina, A. Image annotation and curation in radiology: an overview for machine learning practitioners. *European Radiology Experimental* **2024**, *8*, 11.
67. Wang, J.; Yang, Z.; Hu, X.; Li, L.; Lin, K.; Gan, Z.; Liu, Z.; Liu, C.; Wang, L. Git: A generative image-to-text transformer for vision and language. *arXiv preprint arXiv:2205.14100* **2022**.
68. Yuan, L.; Chen, D.; Chen, Y.L.; Codella, N.; Dai, X.; Gao, J.; Hu, H.; Huang, X.; Li, B.; Li, C.; et al. Florence: A new foundation model for computer vision. arXiv 2021. *arXiv preprint arXiv:2111.11432* **2021**.
69. Li, J.; Li, D.; Xiong, C.; Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Proceedings of the International conference on machine learning. PMLR, 2022, pp. 12888–12900.
70. Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771* **2019**.
71. Kasai, J.; Sakaguchi, K.; Dunagan, L.; Morrison, J.; Bras, R.L.; Choi, Y.; Smith, N.A. Transparent human evaluation for image captioning. *arXiv preprint arXiv:2111.08940* **2021**.