# HEDWIG: Hierarchical Embedding for Deep Weighted Image Geolocation

Jatin Kulkarni
Cornell University
jk2982@cornell.edu

Walter Stark
Cornell University
wzs7@cornell.edu

Madhu Sharma
Cornell University
ms3797@cornell.edu

March 23, 2026

*This manuscript was developed as part of a graduate research project at Cornell University. It was intended for conference submission but was not completed in time and has not been submitted or published.*

## Project Context

This work explores a hierarchical embedding approach for image geolocation. The project was conducted as part of graduate coursework and research efforts, with the goal of developing a conference-quality submission. While promising results were achieved, the work remains incomplete and unpublished.

## Abstract

Image geolocation—predicting the location of a photo—has seen promising advancements through models like PIGEON, which leverage the CLIP model for geographical embeddings. However, CLIP's general-purpose design introduces limitations in location-specific tasks due to its averaging of multi-view inputs. In this paper, we propose HEDWIG, a geolocation model based on ViCLIP, designed to process multi-frame panoramic imagery and textual captions using spatial-temporal attention and a fine-tuned projection layer. Compared to CLIP, HEDWIG reduces the median top-1 prediction error by over 1600 kilometers and increases the proportion of predictions within 750 kilometers by nearly 4×. We also explore scalable captioning strategies using InternVideo2.5 and synthetic panorama generation via Stable Diffusion. Our method significantly improves geolocation accuracy while enabling more robust and efficient inference pipelines for global-scale deployment.

## 1 Introduction

GeoGuessr is an online game in which a user is dropped into a random street through the use of Google Streetview and needs to figure out where they are in the world based on their surroundings. Users are allowed to look around in a 360 degree view as well as move around the street to gather clues about the location before they give their prediction on a world map.

There are previous papers that have implemented algorithms to determine the user's current location based on multiple panoramic images taken from Google Streetview. This is an interesting topic outside of the GeoGuessr game itself, as it can be applied within intelligence and security findings in many different areas.

## 2 Related Work

Diving deeper into the previous papers, this paper is an extension on PIGEON: Predicting Image Geolocations [1]. This paper presents a "new geolocalization system that combines semantic geocell creation, multi-task contrastive pretraining, and a novel loss function" to accurately identify a location on the world map based off of a photo. This new geolocalization system was introduced

through two models, PIGEON and PIGEOTTO. Both of these models perform the same classification task; however, PIGEOTTO is trained on far more images that are scraped off of Wikipedia and Flickr. For this paper, model construction will be based on the PIGEON model to simplify the process in this project.

There are many other papers that inspired the PIGEON model, such as DeepGeo: Photo Localization with Deep Neural Network [2], where they investigate how to improve the accuracy of the results given for more specialized locations.

Collectively, these papers have provided interesting perspectives in improving the location classification problem.

## 2.1 Project Innovation

One of the main innovations used in the paper "PIGEON: Predicting Image Geolocations." is the Contrastive Language-Image Pre-Training (CLIP) model [3]. CLIP is a general-purpose model developed by OpenAI that is trained to connect images with their captions. It is trained on an extremely diverse dataset, enabling it to be versatile when it comes to vision-language tasks.

In the PIGEON paper, CLIP is used to generate geographical embeddings for an image of a Google Streetview that is the input, which is then used to predict the geographical locations in which the image could take place by computing an L2 distance against predefined location clusters computed by an OPTICS model [4].

CLIP alone may not be best suited for this application, particularly because the geoembeddings it generates are averaged across images taken at a location with varying headings (0°, 90°, 180°, 270°). This averaging approach can introduce significant challenges, as the visual information available from different headings at a single location may vary drastically. For instance, one heading might capture a mountain, while another might display an urban road, providing conflicting visual information for geographical estimation. Averaging these geoembeddings could result in embeddings that represent a location entirely different from the actual one, thereby reducing accuracy.

With this in mind, ViCLIP [5] could change the architecture and improve the training process with its specialization in geolocation.

- Instead of taking an average of different geo-embeddings for the different angles of a specific location, ViCLIP uses a weighted aggregation mechanism that can identify geographical features in a photograph such as signs that contribute more to the final embedding compared to ambiguous features like the sky.
- ViCLIP can process a sequence of images through a temporal/spatial transformer which can identify the relationship between multiple photos. When considering the idea of a panorama, a panorama can be sliced into multiple different images that can work well with the ViCLIP model. A generated embedding would then be able to describe all the photos at once while distinguishing their key details.
- Lastly, the ViCLIP model was trained in a way in which embeddings of images from the same region tend to cluster closer together, which is perfect for the image location classification problem this paper aims to improve upon.

Note: Geo-captions are captions generated from an image using an LLM to provide some context relating to its location. Geo-embeddings are embeddings created from a CLIP-based encoder that has a set of images and geo-captions as input.

Outside the ViCLIP aspect, this paper differs from PIGEON in utilizing a new state of the art captioning model for videos, InternVideo2.5 [6]. This decision was made as it was much more effective at captioning a panorama by treating it as a sequence of temporally and spatially coherent frames, thereby enabling richer and more contextually grounded descriptions. InternVideo2.5's advanced understanding of spatial dynamics and scene continuity made it particularly well-suited for panoramic video content, outperforming previous models in both accuracy and descriptive depth.

## 2.2 Result Evaluation

This study evaluates the model using a methodology inspired by the PIGEON and PIGEOTTO models. As the original models are not publicly available, a new model was developed entirely based on the methodologies described in those studies. To assess accuracy, metrics such

as average distance, median distance, various percentiles of distances, and the percentages of predictions within specific distance ranges are utilized.

### 2.3 Project Rationale

This work is relevant as it significantly reduces the computational cost of geolocation models by replacing the heavier parts of the architecture with more accessible and efficient alternatives while improving the results. Through reducing the complexity, users can achieve faster location identification times, which is critical when considering the field of surveillance, emergency response, or navigation. Additionally with the reduced complexity, these models can be deployed on less powerful devices, making it potentially applicable for cell phones, extending the utility to more remote areas.

## 3 Data

In this section, the dataset creation is discussed, as well as how ViCLIP embeddings were generated.

### 3.1 PIGEON Paper Dataset

For the dataset, the PIGEON paper uses 100,000 randomly sampled locations from the game of GeoGuessr, which each consist of 4 different images spanning a "panorama" for a total of 400,000 training images. In addition, the paper uses the Media Placing Eval 2016 dataset (mp_16), Google Landmarks V2 dataset (google_landmarks), and Street View dataset (streetview_cropped).

### 3.2 Initial Analysis - Congressional District Dataset

This study initially focused on experimenting with a specific feature of the PIGEON paper using a simplified yet purposefully constructed dataset in the United States, followed by a world-wide examination. The United States was chosen due to its unique combination of diverse geographical features, uneven population distribution, and well-defined administrative boundaries, making it an ideal dataset for testing the proposed methodology. To achieve

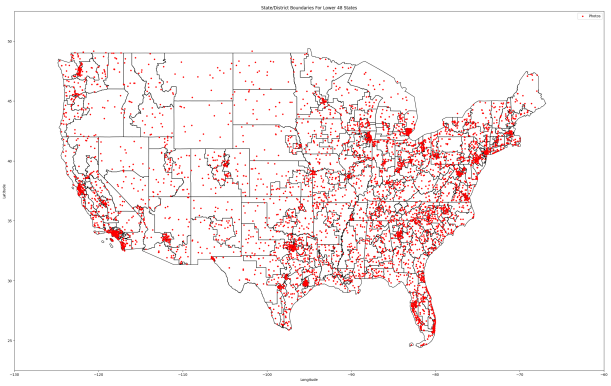

Figure 1: Photo Distribution for Lower 48 States

this, the 435 congressional districts in the United States were identified. Specifically, a dataset consisting of Shapefiles was utilized to incorporate geospatial vector data representing the boundaries of these districts in the form of polygons. By combining GeoPandas with the data contained in these files, the map of the United States was reconstructed with longitude and latitude information, as illustrated in Figure 1.

### 3.3 Initial Analysis - Image Generation

With the new map, Google Street View API was then used to generate photos from random locations within these districts where each district received the same number of photos.

This approach mimics the natural distribution of photos taken in these areas by reflecting the population density, as congressional districts are indicative of regional population size—areas with more districts generally correspond to higher populations. Additionally, the Google Street View API enabled capturing images at various angles, simulating the panoramic experience users would encounter in GeoGuessr. By tailoring the dataset in this way, this ensured a realistic and diverse representation of geographical data while simplifying the overall complexity compared to the original PIGEON dataset. In the photo in Figure 1, it is visible how the photos would increase in areas with high numbers of congressional districts compared to areas with little congressional districts.

Using the mapped congressional districts, approxi-


Figure 2: Geo-caption Example: A cemetery located in a rural area, with cloudy weather, an eastward view, no visible traffic, and some grass and trees.

mately 33,054 images were collected, distributed equally across all districts. Within each district, a random latitude and longitude within its boundaries were selected, and photos were retrieved for eight different bearings. The use of eight bearings was determined by the ViCLIP model's design, which processes eight distinct images. From the total of 33,054 images, around 7,978 unique locations were represented, resulting in only a few locations per district. This number of locations, compared to the 100,000 used in the PIGEON study, was chosen based on proportional representation, as the United States comprises approximately $4\%$ of the global population, providing a suitably scaled dataset to test the ViCLIP hypothesis.

### 3.4 Worldwide Analysis

The data set generated with Google Streetview was a great starting point. Once the hypothesis on the United States was validated, the research had to be tested on a global scale. Using G3: Geolocation via Guidebook Grounding from UCB [7], it provided access to 80,000 panoramas

Figure 3: Example of Input Images

of size 426x300 distributed around the world. On top of this dataset, this paper utilizes the Street View Panoramas Dataset from Kaggle, consisting of 187,777 streetview photospheres from various locations around the world. Each image has a width of 512 pixels and a variable height. This gave over 240,000 images to work with for this research paper spread well around the world.

## 4 Methods

In this section, this paper will outline implementation details of CLIP and ViCLIP, comparing their differences. Additionally, clustering will be discussed with our use of semantic geocells with haversine smoothing.

### 4.1 CLIP

The implementation of the CLIP model in this project focuses on leveraging multi-modal capabilities to generate embeddings that encode both the photo and attached caption data. The ViT-L/14 CLIP model was selected due to its performance in multi-modal tasks. All images are preprocessed - resized and normalized to match the input requirements of the CLIP model. Afterwards, each image is then converted into a batch format to accommodate the model's batch-processing capabilities. Then for each image, a caption is processed using the tokenize function converting a caption data into a format that CLIP can process. Then with a given preprocessed text and image combo, an embedding is obtained by concatenating the image embedding with the text embedding. These embeddings are then averaged out for a specific location for the different bearings received.

Figure 4: Example of Panoramic Image

Figure 5: Example of Image Frames

## 4.2 ViCLIP

The ViCLIP model consists of the following key components:

- **Input** - ViCLIP processes both visual images and textual data, where images are input as video frames and text is provided as associated captions or descriptions. This setup enables a multimodal understanding of the input data. To leverage ViCLIP's ability to process videos rather than just individual images, the input images are preprocessed into a video format. Specifically, multiple images are stitched together into a single panoramic image spanning 360 degrees as seen in figures 3 and 4. From this panorama, frames of size 640×640 pixels are generated every 200 pixels, creating a sequence of 10 frames as seen in Figure 5. These frames are then used as input to ViCLIP as a video, allowing the model to analyze the full context of the image in a sequential manner rather than processing distinct images independently.
- **Projection Layer** - A linear transformation layer aligns visual and textual embeddings into a unified target embedding space. This alignment is fine-tuned to optimize the embeddings for tasks such as clustering and similarity search.
- **Classification Layer** - The final linear layer maps the aligned embeddings from the projection layer into 100 predefined clusters. This step allows the model to effectively categorize the input data for downstream tasks.

Given a location and all its associated image/text embeddings are retrieved, they are concatenated into a single vector of size 1536. Rather than just working with this concatenated vector, the ViCLIP approach uses the projection layer, a linear layer, to effectively integrate information from both the image and the geo-caption. This approach enhances the understanding of the image by focusing on the most important information, rather than relying on simplistic methods like mean or max pooling to combine vectors. The projection reduces the vector size to 768, making calculations more efficient and manageable for downstream clustering tasks. For the clustering task, the classification layer predicts the most likely clustering for the embeddings using cosine similarity.

During training, initial experiments were conducted using both Cross Entropy and Mean Squared Error (MSE) loss to compute the discrepancy between predicted clusters and true cluster labels. Cross Entropy loss was ultimately selected for this phase of training, with plans for more extensive experiments involving different loss functions in the future. Gradients from the chosen loss function were backpropagated to update the weights of the ViCLIP projection and classification layers.

## 4.3 Utilizing Semantic Geocells With Haversine Smoothing

Following the PIGEON framework for geocell creation, the methodology was adapted to create semantic geocells.

1. The G3 panoramic street view dataset was utilized, specifically utilizing the latitude, longitude to retrieve the country name.
2. From there, geographic boundaries at three hierarchical levels (country, admin-1, admin-2) are loaded in.
3. Geocells were initialized at the admin-2 level, and then geocells were balanced to meet the size constraints, ensuring that each cell contains a sufficient yet manageable number of training examples.

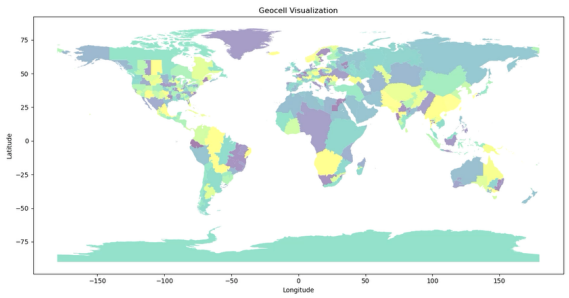


Figure 6: Geocell Visulization for World Dataset

This outputted the photo in Figure 6, where we can see the world dataset into geocells of varying sizes.

For computing the loss, Haversine Smoothing was utilized as a technique to allow predictions near the ground truth to not be punished as heavily as a predication that was completely off.

### 4.4 Coordinate Prediction

This approach generates predictions using a hierarchical approach that combines the model's classification and projection layers. The classification layer first outputs a probability distribution over all clusters. The $k$ highest-probability clusters, with $k = 5$, are identified as the most likely candidates for the geo-location task.

Subsequently, cosine similarity is used to refine the prediction within these candidate clusters. The geo-embedding from the projection layer is compared to every embedding within each of the top $k$ clusters. The embedding with the highest cosine similarity to the geo-embedding is selected, and its associated coordinates are returned as the final prediction.

This hierarchical search method efficiently narrows the search space by combining probabilistic classification with fine-grained similarity matching, enabling accurate geolocation predictions.

## 5 Experiments

For the experimentation in this project, the setup was designed to enable the easy execution of scripts using a YAML file to specify the configurations of the experiments.

### 5.1 Experiment setup

The metrics used to evaluate the results were similar to those in the PIGEON paper, with some modifications made to account for the focus on the United States in this study. The analysis examined the percentage of distances falling within 1, 25, 200, 750, and 2500 kilometers of the actual destination, along with the average distance error rate and a trimmed average distance error. These metrics provided an accurate assessment of the results obtained from the tests conducted.

The following independent variables were identified when working with the US dataset:

1. Captioning - Verbose or Basic. When prompting GPT-4o-mini, a simple description could provide the model with a much easier time of generating a textual embedding compared to a more complex description creating issues.

2. CLIP - Either using ViCLIP or CLIP model. One of the main goals of this paper is to examine how using ViCLIP can potentially improve the accuracy.

3. Pretrained models - Either the model would be trained or not. The question being examined is: Would pretraining the ViCLIP or CLIP model towards the data improve the model's ability to generate better geo-embeddings that could provide a more accurate location?

For global testing, this study focuses solely on comparing CLIP and ViCLIP performance, with modifications to the experimental framework to accommodate the worldwide dataset.

### 5.2 Top-1 Retrieval Distance Comparison

To directly evaluate retrieval quality, we computed top-1 distance metrics for our fine-tuned ViCLIP model (Hedwig) versus OpenAI's CLIP. These include average distance, distance quantiles, and the percentage of predictions below certain thresholds. Results are shown in Table 1.

| Metric | Hedwig | CLIP | Better |
|---|---|---|---|
| Average Top-1 Distance | 7189.75 | 9237.76 | Hedwig |
| Q1 (25$^{th}$ percentile) | 2265.76 | 6819.75 | Hedwig |
| Median | 7413.56 | 9108.90 | Hedwig |
| Q3 (75$^{th}$ percentile) | 10009.99 | 11480.06 | Hedwig |
| % below 1 km | 0.0071% | 0.0071% | Equal |
| % below 25 km | 0.0494% | 0.0141% | Hedwig |
| % below 200 km | 0.7484% | 0.2259% | Hedwig |
| % below 750 km | 8.0138% | 2.0335% | Hedwig |
| % below 2500 km | 26.24% | 10.63% | Hedwig |

Table 1: Top-1 Retrieval Distance Comparison between Hedwig and CLIP. All values in kilometers.

Hedwig outperforms CLIP across all distance quantiles and thresholds. Notably, over of Hedwig's predictions fall within 2500 km, compared to only for CLIP. This supports the claim that the ViCLIP-based approach produces more geographically accurate embeddings, particularly when only the top match is considered.

## 5.3 Ablation studies

To evaluate the contribution of various components in the pipeline, a series of ablation studies were conducted. These experiments systematically removed or altered specific inputs and features to analyze their impact on the final performance. The following ablations were performed:

1. **Panoramic Inputs vs Individual Images** - One of the critical design choices in this project was the use of using Street View to generate individual images for each bearing. Alternatively, panoramic image inputs could also be used to stich multiple frames together to stimulate a broader field of view, potentially helping the model generate richer visual embeddings that better capture geographic context. To evaluate the importance of using individual images for each bearing, the performance with panoramic image frames was compared.

2. **Impact of Captions on Geo-Embedding Quality** - Experiments were not performed without captions due to the nature of our trained projection layer. Since the projection layer is combining visual and textual embeddings, its learned weights the presence of both modalities. If captions were excluded, the text embedding component would default to zero, making the projection layer functionally equivalent to a visual-only model. This would then lead to sub-optimal performance.

3. **Utilizing Stable Diffusion for Panorama Generation** In this work, we leverage ViCLIP to demonstrate its superiority in geolocation estimation. To support this, we utilized Stable Diffusion, introduced by Rombach [8], to generate 360-degree panoramic images from single input images paired with descriptive captions. Specifically, we employed SD-T2I-360PanoImage [9], a text-to-image diffusion model fine-tuned for seamless panoramic generation. This model extends a single image into a 360-degree view using a textual description of the scene.

   Captions were generated using Open AI's GPT-4.0 model with a two-step prompting strategy. The goal was to produce a detailed caption and then condense it to 77 tokens or fewer, adhering to the model's constraints.

   Various prompt techniques were explored, including one-shot captioning and adjusting prompt length, ultimately leading to the two-step approach. This method allows the large language model to generate comprehensive captions and then distill them to 77 tokens, preserving the most critical information for panoramic generation.

   The input image is processed using the SD-T2I-360PanoImage method to produce the resultant panorama, as depicted in Figure 7. This generated panorama can be compared to the ground truth panorama, presented in Figure 8, to evaluate the quality of the output.

   An improvement from the stable diffusion technique is fine tuning. However, the computational cost of fine-tuning large diffusion models has led to the development of parameter efficient fine-tuning (PEFT) techniques. Out of these, Low-Rank Adaptation (LoRA) [10] introduces low-rank updates to the weight matrices of pre-trained models. This will freeze the original paraemters and add trainable low-rank matricies to adapt the model to new tasks.

To adapt the SD-T2I-360PanoImage model for high quality 360-degree panoramas, the G3 dataset was utilized.

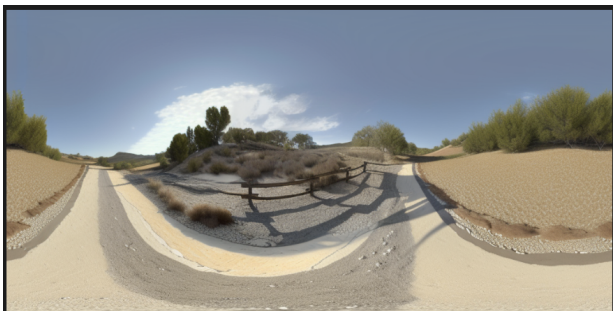

Figure 7: Resultant Panorama generated with SD-T2I-360PanoImage

## 5.4 Pipeline challenges

This process was extremely intensive on compute power, requiring approximately two hours per epoch to fine-tune the ViCLIP model, and then another two hours to generate embeddings from the trained model. The following table outlines in more detail the time it took to perform each step:

| **Step** | **Time Required** |
|---|---|
| Obtained Images | 3 days |
| Generate Captions | 2 days |
| Create Panoramas and Frames | 40 mins |
| Create Train/Test Splits | 20 mins |
| Generate Embeddings - Base Model | 2 hours |
| Create Clusters | 30 mins |
| Train Model | 2 hrs / epoch |
| Generate Embeddings - Trained Model | 2 hrs |
| Create Clusters | 30 mins |
| Testing | 30 mins |

Table 2: Time Requirements for Each Step of the Experiment Pipeline

Figure 8: Ground truth Panorama

For every experiment that was run in this paper, step 4 and onward had to be redone, not enabling enough time to run more epochs for the different experiments.

## 5.5 Results

The expected error of a random guess can be computed to evaluate whether the model provides more value than a random guess as a baseline comparison. The contiguous United States spans approximately 8 million square kilometers and can be approximated as a rectangular region with a width of approximately 4,500 kilometers and a height of approximately 2,500 kilometers. Using a Monte Carlo simulation to calculate the expected distance between random points in this rectangle, the expected value is approximately 1,863 kilometers based on an analysis of 1,000,000 pairs of points.

Table 3: Model Performance Comparison on US Based Dataset: Mean, Median, and Standard Deviation

| Model Used | Caption Type | Base Model/Fine Tuned Model | Mean Distance (km) | Median Distance (km) | Standard Deviation (km) |
|---|---|---|---|---|---|
| CLIP | Basic Captions | Base Model | **1541.55** | **1309.64** | 1038.80 |
| | | Fine Tuned Model | 1780.87 | 1521.91 | 1258.23 |
| | Verbose Captions | Base Model | 1629.50 | 1529.13 | **978.44** |
| | | Fine Tuned Model | 1904.73 | 1709.61 | 1169.87 |
| ViCLIP | Basic Captions | Base Model | 1202.91 | 878.98 | 1091.57 |
| | | Fine Tuned Model | 958.99 | 652.81 | 1034.21 |
| | Verbose Captions | Base Model | 1543.86 | 1468.61 | **851.33** |
| | | Fine Tuned Model | **891.18** | **637.49** | 949.05 |

Table 4: Model Performance Comparison on US Based Dataset: Under Distances

| Model Used | Caption Type | Base Model/Fine Tuned Model | Under 10 km | Under 100 km | Under 500 km | Under 1000 km (%) | Under 5000 km (%) | Under 10000 km (%) |
|---|---|---|---|---|---|---|---|---|
| CLIP | Basic Captions | Base Model | 0.0 | 0.75 | 11.90 | 36.40 | 99.31 | 100 |
| | | Fine Tuned Model | 0.25 | 2.13 | 13.47 | 29.69 | 98.12 | 100 |
| | Verbose Captions | Base Model | 0.12 | 1.00 | 9.46 | 28.57 | 99.18 | 100 |
| | | Fine Tuned Model | 0.0 | 0.25 | 9.83 | 24.18 | 98.99 | 100 |
| ViCLIP | Basic Captions | Base Model | 0.12 | 3.75 | 27.69 | 56.20 | 99.56 | 100 |
| | | Fine Tuned Model | 4.69 | 13.84 | 40.85 | 65.85 | 99.68 | 100 |
| | Verbose Captions | Base Model | 0.06 | 0.93 | 10.02 | 24.87 | 99.24 | 100 |
| | | Fine Tuned Model | 5.38 | 15.35 | 42.66 | 67.10 | 99.87 | 100 |

Overall with the results, an average prediction being off by around 1000 kilometers for a trimmed distance (likely excluding Hawaii and Alaska) showcases that our model is better than a baseline guess on a map. It's also important to consider that this paper was striving to get a working end to end solution first with a smaller data set before developing a more in depth dataset. The next section goes into an error analysis of what this paper will aim to accomplish in the coming weeks.

## 5.6 Experimental Analysis

The evaluation of the CLIP and ViCLIP models highlights the improvements achieved through fine-tuning and the advantages of the ViCLIP architecture. Table 3 shows that fine-tuning reduces mean and median distances for both models, with the ViCLIP fine-tuned model achieving the best performance, including a median distance of 637.49 km with verbose captions. This performance can be attributed to ViCLIP's ability to process more images (8 versus CLIP's 4) and its continuous embedding generation, which avoids the averaging approach used by CLIP. By processing all images and captions together, ViCLIP produces more cohesive embeddings, enabling it to better cluster geospatial data.

Fine-tuning improves the projection layer, allowing embeddings for similar coordinates to cluster more closely while distinguishing disparate locations. Verbose captions further enhance performance by providing richer contextual data. These benefits are most evident in ViCLIP, where fine-tuning results in lower median distances and reduced variability. Table 4 reinforces this, as the ViCLIP fine-tuned model achieves significantly higher percentages of accurate predictions under shorter distances (e.g., 10 km and 100 km) compared to the CLIP model.

ViCLIP's lower median distance and narrower interquartile range (IQR) reflect its consistency and accuracy. Additionally, the CLIP model shows more extreme outliers, with distances exceeding 5000 km, indicating instability in its embedding generation. In contrast, ViCLIP demonstrates fewer and less severe outliers, further highlighting its robustness. The proximity of the mean to the median in the ViCLIP model underscores its stability, while the skewed distribution in the CLIP model

illustrates greater variability.

The decision to base the analysis on median distance, rather than the mean, ensures robustness against outliers such as those introduced by distant locations like Alaska and Hawaii. While the mean and median are generally aligned, the median provides a clearer reflection of typical performance. Both models are expected to benefit from extended training. Due to computational constraints, training was limited to one epoch, but early experiments suggest that additional epochs could significantly reduce loss and improve the quality of geo-embeddings.

In summary, the fine-tuned ViCLIP model demonstrates superior performance due to its ability to process more images jointly and leverage verbose captions effectively. It outperforms the CLIP model in both accuracy and stability, as evidenced by reduced median distances, lower variability, and fewer extreme outliers. Future work should focus on extending training duration and further optimizing caption quality to maximize model performance.

### 5.7 Error Analysis

To improve the project further, the following things were considered to be the most important:

1. **Expand the Dataset**: The primary objective is to expand the dataset by incorporating additional images and locations. This includes using the original randomized coordinate points and supplementing them with images from congressional districts. By utilizing the Google Street View API, a more comprehensive set of images will be gathered to ensure improved coverage across the regions of interest. The current dataset comprises 33,054 images, of which only 7,978 have captions. To address this limitation, captions will be added to the remaining images, and further data will be collected by extending the runtime of the existing data collection process. This effort will significantly increase the dataset's diversity and size, providing a stronger foundation for model training.

2. **Analyze Clustering Methodology**: Finally, this work will experiment with different clustering configurations to optimize the model's ability to identify locations. Using capacity-constrained k-means, this work will test varying numbers of clusters to determine the configuration that yields the most accurate results. Additionally, different hierarchical clustering strategies will be explored, beginning with fewer, larger clusters and progressively refining them into smaller clusters. This staged approach could enhance the model's capability to pinpoint specific coordinate points with greater precision. By evaluating these variations, this papers aims to identify the best clustering strategy to improve geolocation accuracy.

3. **Outliers**: Working with Alaska and Hawaii caused predictions that were off to be an extremely negative impact for the average error distance. For context, Hawaii is roughly 8,500 kilometers away from the furthest location in the United States, which could play a big role in the larger average distance computed. Though working with these two areas introduced interesting challenges with generating images, excluding them within this project could provide a more clear analysis on CLIP versus ViCLIP.

4. **Caption Improvements**: A future experiment that could be conducted is changing the format of captions. This paper analyzes in an experiment the difference between verbose and basic captions, but this could be measured in many different ways. Experimenting with different formats of captions could link unique insights.

## 6 Conclusion

This study highlights the effectiveness of our ViCLIP-based model, Hedwig, for image geolocation tasks. In head-to-head comparisons with CLIP on top-1 retrieval accuracy, Hedwig demonstrated substantial improvements across all quartile and threshold-based metrics. Specifically, Hedwig reduced the average top-1 distance by over 2,000 kilometers and showed better performance in nearly all percentile breakdowns. These findings confirm the model's superior ability to retrieve the most geographically accurate neighbor, a critical trait for applications requiring pinpoint localization.

While promising, several limitations remain, including computational constraints and a dataset focused primarily on the United States, which limits generalization. Fu-

ture work will explore broader geographic coverage, improved multi-frame fusion strategies, and enhanced loss functions to further refine embedding alignment. Additionally, exploring scalable distributed training and inference will be critical to deploying this model in real-world, time-sensitive applications.

# References


[1] L. Haas, M. Skreta, S. Alberti, and C. Finn, "Pigeon: Predicting image geolocations," 2023.

[2] S. Suresh, N. Chodosh, and M. Abello, "Deepgeo: Photo localization with deep neural network," 2018.

[3] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," *CoRR*, vol. abs/2103.00020, 2021.

[4] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, "Optics: ordering points to identify the clustering structure," *SIGMOD Rec.*, vol. 28, p. 49–60, June 1999.

[5] Y. Wang, Y. He, Y. Li, K. Li, J. Yu, X. Ma, X. Chen, Y. Wang, P. Luo, Z. Liu, Y. Wang, L. Wang, and Y. Qiao, "Internvid: A large-scale video-text dataset for multimodal understanding and generation," *arXiv preprint arXiv:2307.06942*, 2023.

[6] Y. Wang, X. Li, Z. Yan, Y. He, J. Yu, X. Zeng, C. Wang, C. Ma, H. Huang, J. Gao, M. Dou, K. Chen, W. Wang, Y. Qiao, Y. Wang, and L. Wang, "Internvideo2.5: Empowering video mllms with long and rich context modeling," 2025.

[7] G. Luo, G. Biamby, T. Darrell, D. Fried, and A. Rohrbach, "G$^3$ : $Geolocation via guidebook grounding,'' Findings of EMNLP$, 2022.

[8] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," 2022.

[9] M. Feng, J. Liu, M. Cui, and X. Xie, "Diffusion360: Seamless 360 degree panoramic image generation based on diffusion models," 2023.

[10] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," 2021.