Project Description

1 Introduction

As of 2023, drones have been recognized as one of the top technologies of the decade by IEEE Spectrum [7]. Their prominence is further evident from the fact that the Federal Aviation Administration (FAA) has registered a staggering 855,860 drones in the United States [1], with the number continuing to grow exponentially. Beyond their traditional military applications for surveillance, drones have evolved into indispensable tools with diverse functionalities, encompassing areas such as environmental monitoring [138], infrastructure inspection [9], and urban development initiatives aimed at creating livable and secure communities [80, 2, 194], to name a few. These functionalities significantly improve the quality of human life. Notable examples of drone applications include the remarkable efficiency of crop spraying, which outperforms manual spraying by 40-60% in terms of speed [45]. Additionally, drones have played a pivotal role in search and rescue operations worldwide, successfully saving over 1,000 lives [84]. Such extensive and diverse utility of drones has been made possible through the integration of artificial intelligence (AI) tools. Looking ahead, the next generation of drones is expected to possess the ability to understand the visual world even more effectively than humans empowering them to make rapid decisions that not only safeguard human lives but also augment laborious human tasks. This feat is attainable by deploying advanced AI algorithms, specifically computer vision.

AI algorithms leveraging deep neural networks (DNNs) have achieved remarkable progress in image and video analysis. However, this advancement is not equally substantial when dealing with images captured from aerial perspectives. Current AI models are typically trained on large-scale images obtained from web sources [92, 163, 102, 130], which often possess weak annotations derived from their accompanying captions. However, aerial images are scarce on the web due to the challenging process of acquiring such data and concerns about data sensitivity [65, 107, 113]. Consequently, the availability of extensive aerial datasets presents a challenge, and recent AI models [93, 66, 20, 198] trained on standard images [116, 61, 101, 148, 92] struggle to generalize effectively to images taken from an aerial viewpoint due to variations in perspective.

Additionally, aerial perspective presents inherent challenges stemming from the presence of small-scale objects against expansive and complex backgrounds, occlusions, and variations in lighting and shadows. These factors vary across different geographies which collectively contribute to the complexity of aerial image analysis. In addition, existing literature reveals a gap in the availability of a comprehensive aerial visual framework in terms of expansive datasets and a robust model that can serve as a foundational backbone for subsequent aerial vision-related tasks. Consequently, significant efforts are required to develop effective and efficient visual models tailored specifically for aerial perspective. This involves addressing issues related to data collection, acquiring discriminative representation learning from limited aerial data, developing efficient models that can be deployed on resource-constrained devices like drones for real-time inference, and ensuring model fairness across different geographic locations.

Main idea. The goal of this research project is to develop an aerial image analysis framework geared toward formulating foundational models, encompassing both vision and language modalities, suitable for a variety of subsequent aerial vision tasks. The key idea is to develop **a geography-aware foundation model for aerial visual analysis** in a hierarchical manner. We envision our foundation model, crafted by aligning language embeddings with aerial image representations, as indispensable for integrating contextual information derived from aerial images, necessitating a geography-aware approach. To encode the geographic nuances within the aerial vision-language embeddings, we propose region-specific local models to facilitate fine-grained aerial visual representation (see Figure 1). Moving beyond the simple ensembling of local models, we introduce the concept of learning a foundation model within a hierarchical tree structure alongside local models to embed the geographic context within the local models. The hierarchical training process will be tailored to specific geographic locations through dedicated optimization techniques design,

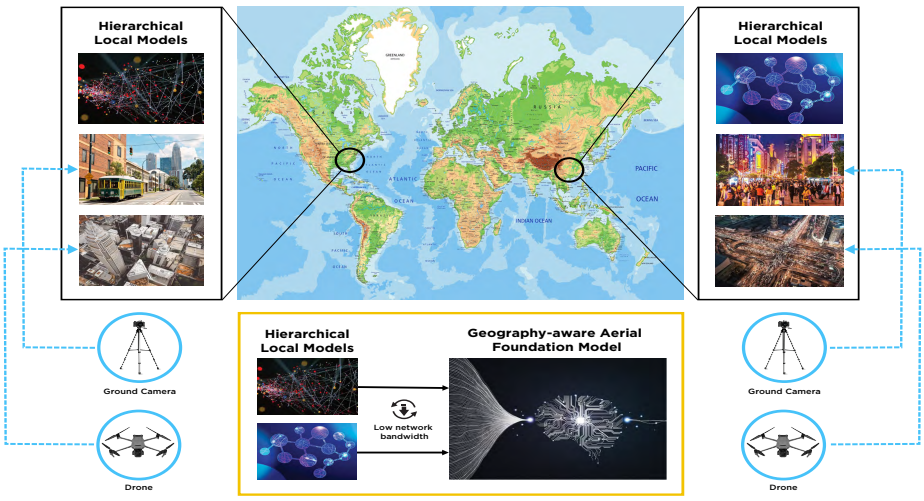

Figure 1: **Geography-aware Aerial Foundation Model.** Our proposed framework consists of multiple local visual models tailored for interpreting aerial viewpoints captured via drones. The local models leverage the visual representation learned from the ground view through viewpoint transfer and language guidance. At a higher level, our global aerial foundation model integrates these local models into a hierarchical tree structure.

integrating privileged modalities originating from aerial data sources. Finally, we plan on developing efficient training modules for transformer-based multimodal models and pushing the scope and boundaries of using multimodal data in federated training to address the notion of privacy, which is limited to a few works, and the potential is largely unexplored. Taken together, the envisioned outcome of this research effort is an efficient, global, geography-aware foundation model capable of accommodating diverse viewpoints and modalities, with the potential for generalization across different geographical regions, while addressing the notion of client data privacy.

Approach overview. We identify new research problems as well as unique solutions to existing problems that arise in a step toward developing a geography-aware foundation model for aerial visual understanding. First, to conduct the research, we will create a new multimodal, multiview drone dataset with ground and aerial videos originating from the same geographic location. Our methodology for learning representations of aerial images integrates two principal components: (1) an investigation towards ground to aerial viewpoint transfer for developing an aerial video conversational model that leverages consistency learning across viewpoints and data augmentation through generating corresponding aerial images given ground images using dual-phase diffusion mechanism, (2) developing a geography-aware aerial (multimodal) foundational model that leverages hierarchical training of local models which are efficient to deploy in resource-constraint devices and encodes contextual information about geographic region of the source aerial data, and at the same time, can address the notion of privacy of the raw data collected from diverse public and private domains. The first component focuses on the development of localized aerial visual models tailored to distinct geographical regions. The final component integrates these localized models to build a global aerial foundational model. In *viewpoint knowledge transfer*, we will explore two complementary research directions. First, we will investigate methodologies to obtain representations from a ground viewpoint and subsequently transfer this knowledge to an aerial perspective. We will conduct this knowledge transfer in a semi-supervised setting, where the ground view model, with limited labeled samples, provides discriminative information to the aerial perspective through a parameterized module. Second, we propose an image-to-image translation-based diffusion model to generate aerial images conditioned on a given ground view image. Although such viewpoint transfer is challenging due to the significant domain gap, the availability of ground-aerial origi-

nating from the same scenes will augment the learning mechanism. Finally, towards developing *aerial video conversational model*, we initially introduce the task of visual question answering in aerial videos. Subsequently, we propose to explore methodologies for localizing a target object within a given video based on a natural language query about the object. These tasks, embedding language within visual contexts, will facilitate a fine-grained understanding of the semantics present in aerial videos, thus laying the groundwork for diverse applications. In *developing a geography-aware hierarchical aerial model*, we propose a hierarchical tree structure with localized models derived from the earlier approaches. This model is designed to capture the statistical heterogeneity present within aerial videos, culminating in a global aerial foundational model robust to geographical variability. Towards developing the hierarchical global model, we will address the challenges associated with devices situated in geographically remote locations that may have significantly higher latency, lower-throughput connections, and resource heterogeneity. To counteract these concerns, we will investigate optimized model architectures to expedite both the training and inference of the local models.

2 Intellectual Merit

This collaborative project entails contribution to representation learning, efficient model deployment, and model fairness for aerial visual perception. To the best of our knowledge, this is the first attempt towards developing a geography-aware aerial foundation model. In contrast to the mainstream AI foundation models, which predominantly relied on abundant weakly labeled data from the web, our approaches address the challenge of scarcity of labeled data in this domain. Also, we argue that a centralized model will ignore the data heterogeneity, resulting in contextually biased models. Unlike traditional large model training schema, we advocate for a hierarchical geography-aware representation paradigm. To pave the way for a transformative era of aerial detection, we will leverage diverse data from multiple viewpoints and modalities, captured across various geographic locations. Subsequently transferring this acquired knowledge to an aerial viewpoint in a hierarchical paradigm offers the vision research community a novel strategy for developing foundational models within data-constrained contexts. We believe that the curated dataset resulting from this project will catalyze further scientific exploration in this domain. The outcome of this project will yield a geography-aware foundational model capable of addressing real-world vision tasks from an aerial perspective.

3 Relevant Work

UAV-based datasets. Over the past decade, there has been a significant increase in UAV-centric video and image datasets, driven by their diverse applications and broad societal implications. Despite this proliferation, no single dataset has emerged as the standard for learning a generalized aerial representation. This gap can be attributed to several characteristics inherent to existing aerial datasets: (i) Many are limited in scale and resolution, such as UAVid [125], DOTA [179], MOR-UAV [128], and AU-AIR [18], with others like UAV123 [132], UAVDT [47], DroneVehicle [166], and BIRDSAI [17] suffering from reduced resolutions; (ii) Datasets exhibit varying flight altitudes, for instance, UAV123 [132] operates between 5 to 25 meters, while MOHR [190] is set at altitudes of 200 meters or above; (iii) Certain tasks, such as object detection, become simplified in some datasets due to skewed object distribution (high density). This distribution often mirrors specific demographic characteristics, as observed in Visdrone [196]; (iv) Extensive datasets like Campus [146] follow scripted scenarios, thereby lacking real-world attributes. Therefore, *this research project aims at collecting a large-scale diverse aerial dataset that can accelerate the current advancement of AI models in this perspective.*

Vision algorithms on aerial data. In aerial perspective, object detection stands as a fundamental task. Object detection methodologies utilized in this domain can be broadly classified into two categories: two-stage and one-stage detectors. The former, exemplified by methods such as RCNN [67], Fast RCNN [66], and Faster RCNN [145], operate by deploying a class-agnostic region proposal module. This is subsequently followed by a process that concurrently regresses object boundaries while classifying them. Conversely,

one-stage detectors, such as SSD [122], Yolo variants [172, 106, 173, 64], and FCOS [168], make direct predictions regarding image pixels as objects, resulting in rapid inference capabilities. Specifically, for drone-based applications, lightweight one-stage object detectors like Yolo-NAS [6] are preferred, given their real-time performance characteristics. It should be underscored that the evolution of object detectors for aerial views is often symbiotic with advancements made for ground-level natural images [182, 183, 176]. The rise of vision transformers [22] has indeed catalyzed changes in object detection architectures [20, 198]. However, there has been a limited focus on tailoring detectors specifically optimized for aerial perspectives, which necessitate addressing their inherent challenges.

Bridging different viewpoints. To leverage the data captured from different viewpoints, a significant body of literature exists, though its focus has not predominantly been on aerial visual representation learning. For instance, joint analysis of time synchronized third-person and first-person videos are widely used in robot learning [111, 153, 156]. However, the ground-aerial perspectives have a large domain gap with different FoV. Thus, learning a latent world representation as in 3D TRL [155] or those focused on camera calibration, might not provide optimal solutions for this challenge. *Thus, this proposal will investigate the methodologies of learning the mapping across these complementary viewpoints in a common semantic space.*

Federated Learning (FL) for data heterogeneity. In traditional centralized training of foundation models, improving the generalization capability of the model is crucial. It is usually induced into the models by the use of data-level regularization [192, 171, 189, 157]. Visual models tailored for aerial images inherently exhibit demographic biases owing to their contextual dependencies on the specific landscapes captured in the images. At the core, in contrast to data center based training [41, 184, 48], FL trains the DNN models on *resource constrained* heterogeneous devices without the data propagation [100, 86]. In the ERM formulation, the popular algorithms for training FL models are FedAvg [129], Local GD [94, 95], local SGD [165, 95, 69], Shifted Local SVRG [69], a few to mention. Many recent works have introduced communication compression in traditional FL formulation [100, 144, 140, 159, 8, 185, 140, 70]. To address personalization, data and device heterogeneity, [161] proposed to learn separate but related models for each participating device. Without data and device heterogeneity, [73, 165, 174, 195, 115] independently proposed local SGD. FedProx [109] is a generalization of FedAvg, SCAFFOLD [89] uses a variance reduction; [16] built a FL system on mobile devices. Recently, [74] addressed personalization and data heterogeneity in FL by a dedicated formulation and by using probabilistic training protocol; [13] used compression techniques on top of this probabilistic communication to reduce the communication bottleneck. *However, we note that addressing data and device heterogeneity remains an important open problem in the FL, and this proposal will investigate it in the form of building a general hierarchical framework trained on massive amount of visual data and training large and sophisticated DNN models.*

4 Proposed Research

4.1 A new Drone Dataset for Multiview Aerial Visual RECognition (MAVREC)

In light of the absence of prior datasets providing multimodal and multiview ground-aerial images [146, 196, 125, 197, 187, 18, 179], we will create a benchmark dataset to address this issue. Our vision is for this new dataset to become a standard reference in the domain of visual aerial image understanding, akin to ImageNet [42] for natural images.

Our methodology involves recording visual scenes using multiple drones equipped with cameras and multiple static ground cameras. The drones will maintain a semi-static position, hovering approximately 25-45 meters above the ground, while the ground cameras capture the same scenes from various viewpoints. This approach enables us to obtain multiple perspectives of the same scene, not only from an aerial viewpoint but also from a ground perspective. To ensure the dataset's diversity and minimize locational bias, we will collect data from different geographical locations across European, American, and Asian landscapes, including mixed pastures during the spring and summer seasons. The demographic variety will empower machine learning practitioners to develop generalized models effective across various landscapes. In ad-

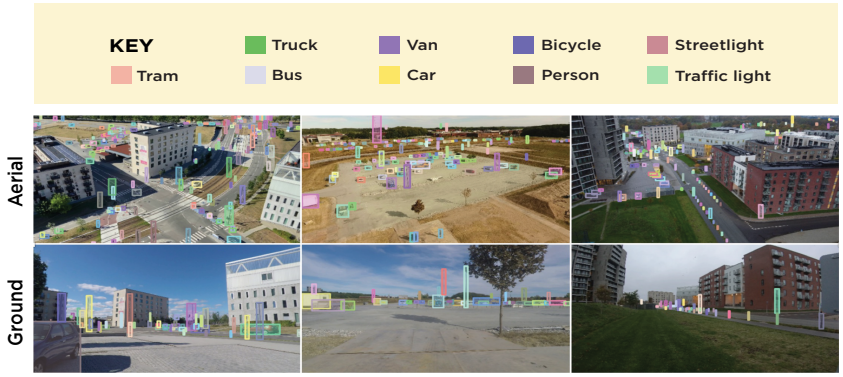

Figure 2: Sample frames from recorded Multiview Aerial Visual RECognition (MAVREC 1.0) dataset.

dition to the multiview RGB visual signals, we plan to incorporate audio descriptions of the scenes. The audio transcripts will provide concise descriptions of the objects, their state changes, and their roles within the scene. We will annotate a subset of frames from multiple viewpoints, specifying the objects of interest along with their boundaries and identifiers in the videos. Moreover, we will have access to text descriptions per frame derived from the audio transcripts. This new drone dataset, we call **MAVREC** for Multiview Aerial Visual RECognition, will facilitate the evaluation of our proposed algorithms in Sections 4.2 - 4.4.

As a preliminary effort, we have already begun data collection [50]. We present sample frames in Figure 2. The current dataset, which we refer to as **MAVREC 1.0** dataset, consists of a dual-view aerial-ground dataset recorded using a drone-mounted camera (DJI Phantom 4, DJI mini 2) and a consumer-grade static ground camera (GoPro Hero 4, GoPro Hero 6, iPhone 11 and 13-Pro) positioned on a tripod. The data was collected outdoors in European settings, comprising 537,000 frames with a resolution of 2700×1520. Additionally, we have manually annotated 22,000 frames for 10 object categories with bounding box annotations, totaling over 1.1 million annotations across both views, averaging 50.01 annotations per frame.

Initial experiments conducted on MAVREC 1.0, as elaborated in subsequent sections, indicate that *the inclusion of ground-aerial videos from a specific geographic location substantially enhances the performance of high-level vision tasks*, although *temporal synchronization between these videos appears non-essential*. This inference, derived from the analysis of the current dataset, suggests that the expanded version of MAVREC may not necessitate time-synchronized aerial-ground footage. Consequently, this realization streamlines and pragmatizes the data collection process. The expansion plan for MAVREC 1.0 involves the acquisition of independent aerial footage via drones from two locations: the University of North Carolina at Charlotte (UNC Charlotte) and the University of Central Florida (UCF) in the USA. This will be complemented by the integration of existing ground-level videos from corresponding geographical areas. We note that no personal data or offensive content is/will be included in this dataset, thus exempting it from IRB (for North America) or GDPR (for Europe) compliance. We have consulted with legal experts to ensure this aspect's thorough validation. The dataset will be made publicly available for research purposes.

4.2 Ground to Aerial Viewpoint Knowledge Transfer

In this study, we aim to investigate the effects of jointly acquired aerial and ground data for fine-grained visual representation learning in the aerial perspective. Given that the aerial viewpoint presents a challenging perspective due to its limited data availability on the web, our initial focus will be on learning visual representation of aerial images with scarce annotated data, while capitalizing on the advantages of ground view images. Subsequently, we will explore the generative models to augment the aerial visual models by generating images of cross viewpoints based on a given perspective. Finally, these novel mechanisms/components can be integrated into a video conversational model to adapt it for the interpretation of aerial videos.

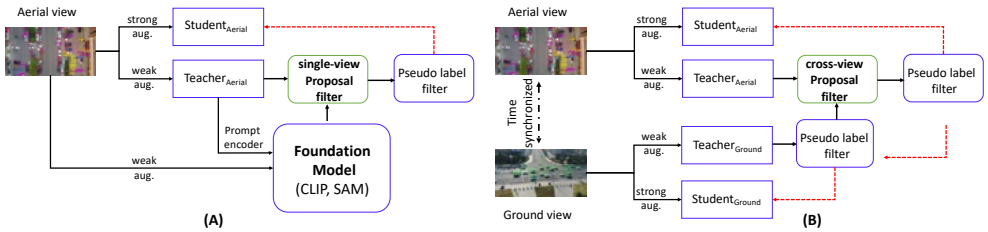

Figure 3: **Teacher-student Framework for Semi-supervised Object Detection.** (A) Demonstrates the process in an aerial view where the Foundation Model is employed to refine predictions by filtering noisy outputs of the aerial teacher. (B) Illustrates the approach for time-synchronized ground-aerial viewpoints, leveraging representations learned at the ground view to similarly filter the teacher's predictions.

4.2.1 Cross-view Consistency Learning for Aerial Object Detection

Background and motivation. The advancement of visual understanding in natural images has witnessed significant progress owing to ongoing research in foundational vision-language models [93, 158, 186, 118, 10, 85, 124]. However, these models suffer from inherent biases stemming from data captured predominantly from a limited viewpoint, thereby hindering their generalizability to videos taken with drones, which are notably underrepresented on the web [71, 96, 91]. Aerial data collection is complicated due to UAV-flight regulations and safety protocols, atmospheric turbulence, and many more [87]. And if collected, it is expensive to annotate (for example, the object boundaries) due to the dense information present in the scenes. Consequently, the development of effective models for aerial visual representation encounters substantial challenges due to the scarcity of annotated aerial images [105]. This motivates us to explore and answer questions such as: *Can we use the multiview property of our proposed dataset to mitigate the challenge of requiring an abundance of annotated data? Can we leverage the models that operate on popular viewpoints (like ground view) and translate that information to the aerial viewpoint?*

Main idea. In this task, we aim to investigate the research direction of learning from limited annotated data. Towards label-efficient computer vision, semi-supervised algorithms [175, 180] have demonstrated considerable progress in recent years by training models on a fraction of labeled data and mostly on unlabeled data. In particular, one successful pipeline in this direction is consistency learning [162, 14], wherein a model is initially trained on the available labeled data. Subsequently, the model is employed to generate pseudo-labels for the unlabeled data, and retraining it on both labeled and pseudo-labeled data. In this task, we aim to investigate the utilization of the multiview characteristic in our proposed dataset to enhance the aerial representation of models. The core idea of this research is to take advantage of the discriminative representation captured by the models when trained on ground view data and subsequently devise a mapping to transfer this knowledge to the aerial viewpoint. It is important to note that both viewpoints may not share the same field of view and could contain dissimilar objects in the scene. Nonetheless, there will undoubtedly exist common objects in both views, for which interpreting the ground view appearance proves to be comparatively easier due to factors like larger object size, reduced occlusion, and availability of abundant captioned data in this view. Therefore, we propose a localization-agnostic module capable of learning the mapping from one view to another through a shared embedding, thus enabling the models to effectively leverage the knowledge gained from ground view data and improve their aerial representation.

Approach. For the proof-of-concept, we will explore two directions aimed at enhancing vision tasks, particularly object detection. These directions involve (i) utilizing pseudo-labels generated from foundational models and (ii) leveraging ground-view object detectors to improve aerial visual representation learning.

In (i), we leverage foundation models like ViLD [71, 96] and SAM [98] to provide robust pseudo-labels for aerial images. However, preliminary experiments reveal that these foundation models are less effective on the aerial viewpoint. To address this limitation, we adopt a traditional teacher-student network, as in [175, 180], within the framework of semi-supervised learning. The teacher and the student are identical object detectors initialized with weights from the model trained on labeled images, and the teacher's parameters

are an exponential moving average of the student's parameters. Pseudo-labels are generated by the teacher, which is fed with weakly augmented aerial images. However, these pseudo-labels are considerably noisy due to challenges specific to object detection on aerial images, such as smaller spatial resolution of objects, occlusion, and variance in object distribution within the scene. We hypothesize that many of these low-confidence object proposals may actually be correct, but they are discarded due to their low confidence. To address this, we employ a foundation model to obtain either a pixel prompt or a textual prompt. This prompt describes the region and object class in text in the image or provides pixel centers of objects. Subsequently, relevant object proposals are obtained by feeding the aerial image to the foundation model, guided by the prompt generated from the noisy teacher. Finally, we utilize a proposal filter that takes into account proposals from both sources (P_A and P_G) and discards irrelevant object proposals, generating refined pseudo-labels for the student.

In (ii), we will train collaborative teacher-student frameworks for two perspectives: aerial and ground. We hypothesize that detecting objects in the ground view is easier than in the aerial view. Thus, we can use pseudo-labels generated by the ground perspective teacher to guide the noisy aerial perspective teacher. To achieve this, we will input object proposals from both teachers, denoted as P_A and P_G, respectively (where $P_A \geq P_G$), into a cross-view **proposal filter**. The proposal filter takes input sets of proposals (P_A and P_G) characterized by their coordinates. These proposals corresponding to different spatial regions are reshaped into fixed-sized features using ROI pooling [145], resulting in Z_A and Z_G. We will explore the concept of an affinity matrix $\mathcal{A}$ defined as $\mathcal{A} = Z_A Z_G^T$, which will learn the correspondence between the object proposals from different input perspectives, either from the foundation model or another viewpoint. The filtered proposals will be obtained by performing differential top-k selection [139, 82] on the affinity matrix.

Preliminary Results. PI Das and PI Dutta initiated this research study as a preliminary proof-of-concept and reported in [50]. The experimental analysis is conducted using the MAVREC 1.0 dataset, which was already collected for this purpose. In the initial results, a semi-supervised framework, Omni-Detr [175] is tested separately on individual aerial and ground perspectives. The average precision achieved for object detection is 19.8% for aerial perspective and 45.8% for ground perspective. These results support our hypothesis that object detection in ground view is easier than in aerial perspectives. Additionally, we explored a naive joint training approach with both aerial and ground view images by sampling their images within the same batch. This joint training strategy resulted in an improved average precision of 26.7% for aerial images. Comparatively, pretraining the model with ground view images and fine-tuning with aerial images yielded inferior results. These preliminary observations demonstrate the potential of our proposed research study to leverage ground view data for enhancing aerial visual representation.

4.2.2 Aerial viewpoint Generation with Generative Models

Background and motivation. In this task, our objective is to generate synthetic aerial images from provided ground view images. These models will enhance the training of our local aerial visual models, given that ground view images from various geographic location are easily accessible in the web. Recent developments in generative models have seen a paradigm shift from GANs [68, 90, 142] towards diffusion models [76, 147, 150]. These diffusion frameworks have facilitated cutting-edge image synthesis. With the advancements in vision-language models, text-to-image diffusion models have shown to generate realistic images from arbitrary text captions. Such models can capture visual concepts, including the interactions between objects, geometry [143, 147, 103, 135, 97]. However, a limitation lies in the minimal control they offer over the resultant content. Concurrently, the domain of image-to-image translation, which determines a transformation of an image from a source domain to a target, maintaining domain-agnostic attributes, is also increasingly influenced by diffusion models [143, 150, 149]. *Despite their potential, diffusion models remain largely unexplored for synthesizing novel viewpoints from a given set of 2D scene images*. Although the generation of new viewpoints has been explored with 3D scenes [177], obtaining such intricate 3D data is often infeasible in real-world contexts.

Main idea. In this task, we will investigate image-to-image diffusion models to generate a novel aerial

viewpoint from a given ground image. The problem is formulated as follows: Given a dataset of input-output image pairs, denoted as $\mathcal{D} = \{x_i, y_i\}_{i=1}^{N}$, which represents samples from an unknown conditional distribution $p(y|x)$. This mapping is one-to-one, ensuring the target image aligns with the input source image (as we utilize time-synchronized images). Our goal is to learn a parametric representation of $p(y|x)$ translates a source ground image x into a target aerial image y. To tackle this, we adapt the denoising diffusion probabilistic model (DDPM) [76] and masked generative transformer [21] for conditional image generation.

Approach. In this task, we will explore two approaches towards image-to-image translation for the generation of aerial images for given ground images. (i) The first approach will employ a conditional DDPM model that generates a target y_0 over multiple time steps. Initiated with a purely noise-influenced image, represented as $y_T \sim \mathcal{N}(0, I)$, the model iteratively refines the image through successive iterations $(y_{T-1}, y_{T-2}, \ldots, y_0)$ according to learned conditional transition distributions $p_\theta(y_{t-1}|y_t, x)$ such that $y_0 \sim p(y|x)$. The intermediate image distributions during inference are characterized using a forward diffusion process. This process is responsible for the systematic addition of Gaussian noise to the image via a predetermined Markov chain. The goal of these models will be to reverse the Gaussian diffusion process by iteratively recovering signal from noise through a reverse Markov chain conditioned on x. Our investigation will revolve around how to learn the reverse chain using a neural denoising model that takes as input a ground view image and a noisy target image and estimates the noise. (ii) Secondly, we will explore text guided masked generative transformer which are more efficient than diffusion and autoregressive models. A naive approach of this framework will be adopted from [21] which will consist of three branches - pre-trained text encoder, base ground model, and aerial visual model. The pre-trained text encoder facilitates text embedding, which then undergoes cross-attention with image tokens across both the ground and aerial transformer layers. The ground base model will use a VQ Tokenizer that is pre-trained on ground images and generates a latent space of tokens. This sequence will be masked at a variable rate per sample and then the cross-entropy loss learns to predict the masked aerial image tokens. Once the base model is trained, the reconstructed ground image tokens and text tokens will be passed into the aerial visual model that then will be learning to predict masked tokens for the aerial viewpoint. Given that the primary ground model offers tokens that correspond to the ground view's latent map, our aerial visual model's operation is refined to translate the ground latent map into the aerial equivalent. This process will combine with the aerial VQGAN decoding the final aerial image representation.

Preliminary Results. To ascertain the necessity of time-synchronized aerial-ground view images, the PIs Das and Dutta conducted experiments on the MAVREC 1.0 dataset. They focused on object detection, for which the D-Detr model was employed. We find an object detector trained from scratch on MAVREC 1.0 achieves mAP of 13.1%. We observe that pre-training the model on a benchmark dataset, Visdrone [196] leads to a 78.6% improvement in object detection performance on the MAVREC 1.0 dataset. Remarkably, leveraging ground view images from the MAVREC 1.0 for pre-training the model manifested a superior performance boost, showing an increase of 129%. This shows the importance of ground-aerial data captured from the same geographic location for pre-training aerial visual models. Consequently, the generation of high-fidelity, time-synchronized aerial-ground imagery is anticipated to significantly augment the efficacy of such aerial visual models.

4.2.3 Ground to Aerial Video Conversational Model

Background and motivation. Recent advancements in Vision Language Models [93, 38, 10, 118, 124] (VLMs) and Large Language Models [136, 169, 24, 72] (LLMs) have accelerated research in image and video grounding through textual transcription. This progress has led to the development of video-language models, significantly enhancing video understanding tasks. Novel methods, including LLaVA [120, 119], VideoChat [108], Video-LLaMA [191], and Video-ChatGPT [133], have emerged, capable of summarizing video content into text and responding to high-level textual queries related to the input video. However, a notable limitation of these VLMs is their inherent bias toward ground-view image understanding, a conse-

quence of training datasets predominantly comprising ground-view images. This bias results in a generalized, coarse-level interpretation of ground-view scenes. In contrast, aerial video grounding demands a more fine-grained analysis and understanding of object dynamics and states of motion within the scene.

Main idea. The objective of this task is to develop an interactive aerial video conversational model, representing an advancement towards an aerial foundation model. This proposed model will be designed to align the aerial video representations with LLMs and ground-view image interpretations. This model aims to address the specific challenges associated with processing and interpreting aerial video data. In contrast to existing video-language models, our proposed model focuses on the detailed analysis of aerial video, providing intricate textual descriptions and object tracking aligned with specific text queries. We envision that such an aerial video conversational model will offer a user-friendly interface for interpreting, searching, and monitoring objects within a scene, incorporating human interaction in the loop. To the best of our knowledge, this will be the first model of its kind, specifically designed for aerial viewpoints. It is uniquely capable of generating object tracklets in response to input queries related to the scene, distinguishing it as a novel conversational model.

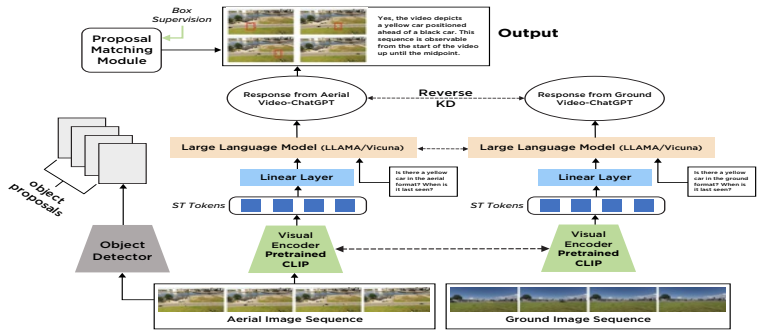

Figure 4: A Ground-to-Aerial Video Conversational Model Framework, featuring a Vision Language Model (VLM), linear embeddings, a Large Language Model (LLM), and an object detector. This framework processes both video and text queries, outputting responses to these queries as well as object tracks pertinent to the LLM's response.

Approach. Our proposed video conversational framework, while conceptually similar to existing frameworks utilizing LLMs and VLMs like CLIP as visual encoders, diverges in key aspects. Firstly, it focuses on extracting optimal spatio-temporal representations from CLIP to capture fine-grained details crucial for aerial video representation. Secondly, our framework focuses on transferring knowledge from ground to aerial videos. Lastly, it aims to output relevant object track proposals by aligning them with the LLM outputs in a shared semantic space. In this task, the aerial video conversational framework is proposed to identify and highlight objects in aerial scenes in response to textual prompts. The development of this model encompasses several key components: (i) leveraging a pre-trained VLM for initial video processing and generation of pseudo-text representations, (ii) employing LLM to interpret these pseudo-text representations alongside input text queries, (iii) implementing a knowledge transfer mechanism to adapt insights from ground-level viewpoints to aerial perspectives, and (iv) integrating an object detection and an alignment module to accurately associate relevant objects within the scene with the corresponding text queries. Our proposed framework broadly comprises two branches, each equipped with a VLM, a linear projection layer, and a LLM to generate text responses corresponding to input video and text queries. These branches are designed to process aerial and ground videos independently, yet they share the VLM and LLM components. Additionally, the aerial videos are fed into an object detector to generate object proposals. These proposals, along with the projected spatio-temporal representation, are aligned in a common semantic space for the extraction of relevant proposals, which constitutes another key output of our framework.

For *utilizing pre-trained VLMs* to map input videos into latent representations for LLMs, we plan to investigate various strategies for pooling spatial and temporal information. These latent representations should capture critical regions within scenes for fine-grained analysis while encoding temporal relationships among frames. Consequently, we will explore an attribute-aware network that disentangles key attributes from the

CLIP [93] representation and employ a graph neural network for their spatiotemporal modeling. Regarding *knowledge transfer between ground and aerial video representations* in language space, our approach unlike traditional sequence-level distillation will explore reverse knowledge distillation [72]. This will involve encouraging the aerial branch to generate text samples that are preferred by the ground branch, a more feasible approach. Finally, in this task, we aim to introduce novel methods for *aligning object proposals with the latent video representation from VLMs* to filter relevant proposals efficiently. The object detector will be trained using ground-truth box supervision, complemented by a self-supervised contrastive loss for object-text alignment.

Preliminary Results. PI Das has initiated the preliminary implementation of the proposed VLM representation for action detection in long, untrimmed videos [30]. In the current setup, a frozen CLIP encoder is employed to extract object semantics, bypassing the need for prompt engineering. A graph reasoning block is employed to model the temporal relationship among the extracted attributes (objects in the scene). The efficacy of this implementation has been tested on the Charades dataset [160]. Using only the extracted CLIP features as a baseline, the system achieved a mean Average Precision (mAP) of 18.1%. In contrast, our attribute-aware network without prompt engineering—demonstrated a significant improvement, yielding a 31% mAP. This shows the potential of using such a concept for fine-grained video representation in our aerial video conversational framework.

4.3 Developing a Geography-aware Hierarchical Aerial Visual Model

In this study, we plan to explore how to learn hierarchical model architectures systematically and collaboratively from *heterogeneous data* that are coming from different geographies. We envision, at every sub-level of this hierarchical learning, resultant models would serve as a *geography-aware aerial visual basis sets* to perform efficient inference of any new dataset that shows feature similarity to any particular geographies; see Figure 1. This hierarchical approach would build models that would learn from the statistical heterogeneity of the data, and its completion would lead to a *global aerial foundation model* that is robust to diverse geographies and has superior *inter-domain inference quality* than the recent ensemble-based or averaging-based approaches [178, 44, 137, 81, 83]. Subsequently, training geographically remote large DNN models performing sophisticated technical tasks such as VQA or aerial view-point generation requires computational efficiency. Therefore, we also focus on proposing reduced computational protocols while training these models.

4.3.1 Towards a Global Aerial Foundation Model from heterogeneous training data

Background and motivations. We witnessed a surge in open-source UAV-based video and image datasets collected across different geographies (primarily in Asia and North America) in the last decade. These datasets have an intrinsic unexplored property regarding their *scene color content*. Many studies in social sciences, humanities, and natural sciences reveal latitude influences the population density and, hence, the color content of the scenes. Moreover, the ambient light level and variation of outdoor illumination is a function of solar elevation and depends on multiple factors [164]. These seemingly low-key factors, in addition to the geographical differences of the regions, make the datasets naturally diverse and create contextual heterogeneity of the scenes collected in different pastures of the same town; see Figure 5 as a result of our initial investigation. Consequently, the inter-domain inference quality of the DNN models trained on these heterogeneous data vastly differ. E.g., our *preliminary investigation* shows that state-of-the-art object detection models, including the *open-world foundational models* like Grounding-Dino [121], which fail to achieve the expected performance level on MAVREC. This observation validates an inherent bias of these models towards ground-view data. Moreover, an object detector pre-trained on popular ground-view dataset (MS-COCO [116]) or other aerial datasets collected from different geographies (e.g., Visdrone [196] from China) has diminished efficacy on aerial images obtained from disparate geographical regions (for our case, Europe). Therefore, unlike classical object detection, training a sophisticated DNN model on a large dataset (e.g., ImageNet [148] or MS-COCO [116]) does not offer the best overall solution. We find augmenting ob-

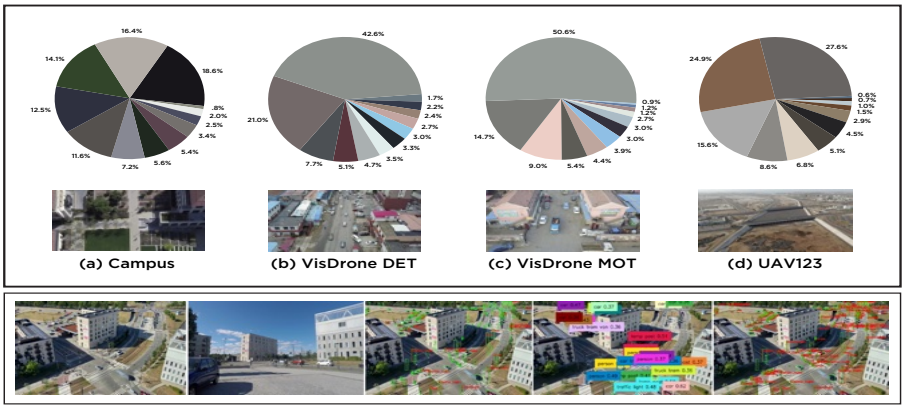

Figure 5: **Top row:** Most dominant colors in the sample frames of (*a*) the Campus dataset [146], (*b*) VisDrone DET, (*c*) VisDrone MOT [196], and (*d*) the UAV123 dataset [132]. **Bottom row (L to R):** Sample frames from time-synchronized MAVREC showing aerial and ground views; D-DETR [198] trained on aerial VisDrone DET [196] inference results on MAVREC (GT bounding boxes are green, and detection are in red); Grouding-Dino [121] inference result on MAVERC; inference results of D-DETR trained on aerial MAVREC has fewer missed detections.

ject detectors with ground-view images from the corresponding geographical context is a superior strategy that boosts detection performance.

In the machine learning (ML) community, *model soups* (combined representation of DNN model parameters of multiple networks) [178], *model ensembles* (combined outputs of many DNN models) [44, 137], *stochastic weight averaging* (combination between ensembling and averaging model weights) [81] are conventional approaches for fine-tuning a large pre-trained DNN model and maximizing its accuracy. These approaches have benefits and drawbacks over each other in performance, calibration, applicability, and computational time—no technique is provably resilient to the diversity of the heterogeneous data; one size does not fit all. Recently, population parameter averaging (PPA) [83] emerged as an approach that "combines the generality of ensembling with the efficiency of weight averaging" and is an efficient way to debunk the computation cost of model ensembles — in a nutshell, claims to be the best of both worlds. Taken together, the color diversity analysis of different state-of-the-art drone-based datasets in our initial investigation, and the remarkable success of ensemble-based models in the ML literature, lead us to ask the following questions: (*i*) Data heterogeneity is a broad and nuanced term, and many DNN training protocols are agnostic to the peculiarities of the data heterogeneity. Can we design a *geography-aware aerial visual basis sets of models* in a hierarchical manner where the local models would learn from the statistical heterogeneity of the data and eventually lead to a *global foundation model*? (*ii*) An expert in ML would commonly agree that PPA is a form of federated training and opens a plethora of opportunities and challenges from theoretical and practical points of view. Can we incorporate many of these ideas in our hierarchical training protocol and perform a *nontrivial* study, both theoretically and empirically?

Main idea. In this task, we envision building a general hierarchical model tree where the local models would learn systematically and collaboratively from *heterogeneous data*. Let $\mathcal{D}_{i,1}$ be *heterogeneous dataset* collected from n different *sub-local regions*; see Figure 1. E.g., In Orlando and its surroundings, UCF has several campuses; Orlando is a sub-local region. Among them, the Orlando downtown campus is in a prime urban location (say, $R_{1,1}$), UCF's main campus is in a mixed urban pasture with some greeneries in the surrounding ($R_{2,1}$), UCF's Lake Nonah campus is a newly developed architecture with an overabundance of flora in the region ($R_{3,1}$), and UCF's Cocoa campus has a prime water-body (Indian river) and mangrove flora ($R_{4,1}$). Therefore, we can train DNN models (with parameters $\{x_{i,1}\}_{i=1}^{4}$) *collaboratively* on the data curated at these four sub-locations (with various data augmentation and pre-training strategies) and obtain a refined sub-local model ($\bar{x}_{1,1}$) that we postulate would represent the Orlando area. It could be possible that some of these places have limited network bandwidths and/or limited computing resources, and we

can address them in our general framework. Nevertheless, we train several local models from p different regions and create the base nodes ($\{\bar{x}_{i,1}\}_{i=1}^{p}$) of our hierarchical model tree, for level $j = 1$. We can train and fine-tune these local models (one hierarchy above the fine-grained base models; note that, smaller j represents more regional fine-grained model) collaboratively, and this process continues until we reach a central global foundation model ($\bar{x}$, with $j = L$, the total number of levels). At each level, j, of this *tree-like hierarchical structure*, we hypothesize the collection of the local branch models conjointly acting as a *basis set*, denoted as $\mathcal{S}_{i,j} := \{\bar{x}_{i,j}\}$. During inference of an unknown dataset, if the user has abundant time and resources available for feature extraction and feature projection of the new data to a trained local basis set, $\mathcal{S}_{i,j}$, of a particular level, j, then a weighted combination, $\sum_{i=1}^{n} w_{i,j}\bar{x}_{i,j}$, with weights, $w_{i,j} \geq 0$ of the basis models, $\bar{x}_{i,j}$ would lead to a better inference. The depth of the level j depends on the resource constrain and is need-specific. However if time and resources are limited, the user can use the *global foundation model* for fast inference without compromising too much in accuracy.

Approach. By building a hierarchical tree structure with local models, we want to propose a *general framework* that would contain federated training as a particular case (one can think of each node, i, of this tree for a level j as a federated training paradigm that trains local models from a particular region); see Figure 1. However, it is not exclusively a federated training paradigm; in some local regions, the training can happen within a data center with a central server and worker nodes connected by fast network bandwidth; in some other regions, training can happen in a decentralized manner. Since our initial result indicates complex data heterogeneity, for optimization training algorithm at each node, we plan on designing new communication-efficient local training algorithms robust to data-heterogeneity and personalization (note that, popular FedAvg [129] can diverge on highly non-identical data partitions; addressing data-heterogeneity remains an important open problem in FL). Designing efficient training algorithms is both a theoretically and empirically challenging endeavor—we will train large DNN models for sophisticated technical tasks such as VQA or aerial view-point generation. Moreover, federated training has a rich footprint of different protocols to aid in personalization [78, 114], privacy, and data heterogeneity, namely, client selection (e.g., Lazy gradient aggregation [23], power-of-choice selection [26], clustering [123], etc.), decentralized training [99, 112], and compressed communication [48, 185, 184, 151]; we plan on incorporate them for efficient training and explore them theoretically. To the best of our knowledge, except for a few works [185], the ML community performs federated training in learning smaller DNN architectures on small-scale datasets (MNIST, F-MNIST, CIFAR-10) in *artificially infused virtual training environment.* We plan to explore real training scenarios where the clients have real geographical remote locations, and we plan to train large encoder-decoder-based architectures and experience the empirical resilience of our proposed model.

4.3.2 Hierarchical Training Meets Multimodality

Background and motivation. The aim and scope of FL is to maintain the client's data privacy and consequently introduce *personalization* to the client's model learning based on data heterogeneity that has been achieved by designing a dedicated optimization algorithm [109, 13] or by architecture design [141]. However, in the transformative era of large foundation models in diverse digital modalities (language, vision, and multimodal), the de facto training protocol requires a huge corpus of high-quality data. Although data abundance, at present, is not an issue, privacy concerns and stringent regulations of various data domains, shortly, constrain the raw data sharing that is necessary to train these models—Experts predict, "We may run out of data to train these models by 2025 [4, 3, 5]." Hence, the obvious question is that if the data sources dry out, can we solve the issue of training large multimodal foundational models in a modality-collaborative environment?

Modality collaboration, or *multimodal fusion*, allows different modalities to train and learn collaboratively by sharing high-level common knowledge and has recently been introduced in federated learning [62, 181, 193]. The scope of multimodality in FL is limited to a handful of works, and the potential is largely unexplored. E.g., earlier works consider a homogeneous model for each modality [181, 193]; only recently, [25] took a step towards a larger server model training by knowledge transfer of diverse smaller client mod-

els, and [188] use knowledge-transfer between uni and multimodal clients with heterogeneous data and learn a larger global model. It is worth noting that some multimodal frameworks are restrictive, barring the participation of unimodal clients [181]; [188] lifts this barrier but still relies on multimodal clients. The proliferation of data heterogeneity and modality in our previously described approaches (see Figure 4) and potential privacy concerns in many video/image data domains motivate us to propose a multimodal learning paradigm in some branches of our hierarchical training process.

Main idea and Approach. Benchmarking on **MAVREC 1.0** indicates a greater aerial visual perception when ground images are used in the pretraining strategy. Interestingly, these images do not have to be time-synchronized with the aerial view, and no sequential information is required. This opens the possibility of augmenting MAVREC with multiple datasets, and at the same time makes us rethink the notion of privacy for some data sources if there is such a requirement. The privacy issue, along with the research questions depicted in the previous Sections that require multimodality of the data and hierarchical model learning paradigm, we can *nontrivially* invoke multimodal FL as one of our main training approaches in some *sub-branches of our hierarchical aerial visual model* where privacy is an issue. For instance, referring to the hierarchical paradigm in Figure 1, we can imagine that some geographical sub-locations have private multi and unimodal training data and they can participate in a modality collaborative private training. We can answer several questions from both practical and theoretical points of view: (*A*) Presently, practitioners use different federated training algorithms, such as FedAvg [129], Scaffold [89], etc., and combine them with uni and multimodal approaches, as highlighted before. To the best of our knowledge, none of these algorithms give any theoretical convergence guarantee for multimodal federated training, nor do they explore modality-dependent personalization at different clients. Moreover, the data sources used for this multimodal training are small-scale. On top of that, we can push the limit of transformer-based models used in our different learning tasks which presumably can tackle the multimodality of the data due to the attention map. This setting has only been explored under extremely limited settings in unimodal FL [141]. We envision that we can lead the ML community in better directions in solving the above-mentioned problems, with high-quality digital data, as in MAVREC, large multimodal learning models, combined with the complex hierarchical training architecture. (*B*) In the multi-modal setting, generally, a modality with more training samples tends to produce better generalizability. Presently, this is a practical observation in a limited setting. We want to explore this both in the real-world setting and explain theoretically why it is the case. Similarly, imbalanced numbers of samples lead to failure. Will our hierarchical model be resilient to this imbalance? Finally, the idea of parameter sharing is another open problem in the present multimodal FL setting.

4.3.3 Efficient Learning Modules

Background and motivation. In our hierarchical model tree, geographically remotely located devices may have significantly higher latency, lower throughput connections, and resource heterogeneity. Performing the training via compressed gradient communication [184, 185, 77, 32, 11, 48, 151] is one way to mitigate this problem. However, we plan to work on complex encode-decoder architectures such as Transformers. Therefore, we will achieve additional benefits if we can address the computational overhead of the Transformer's core architecture. Kernel methods (used primarily in Transformer's encoder architecture) project data points into a high-dimensional space and find the optimal splitting hyperplane in that space. In the linear kernel method (that is, the inner product used in Transformer's encoder's attention module), the data, $X \in \mathbb{R}^{n \times d}$ is represented in the Gram matrix, $\hat{K}$. Calculating the matrix $\hat{K}$ becomes computationally expensive when the number of points, n, is too large. Large n is one of the reasons that contribute to the success, and at the same time, poses humongous computational overhead to the encode-decoder-based Transformer architectures. In the ML literature, efforts have been made to reduce the effect of large n [167]. However, *can we reduce both the input numbers, n and feature dimension, d, maybe conjointly?*

Main idea and Approach. We will briefly outline our strategies: (*a*) **Sparsification of the features.** DNN training often correspond to *sparse* tensors, and most of the DNNs are over-parameterized [170]. Sparse tensors are often direct artifacts of the training process; e.g., the gradients of the NCF [75] and DeepLight

[43] models consist of roughly 40% and 99% zero elements, respectively. Furthermore, one popular lossy sparsification strategy, Top-k, with k as little as 0.1% of the gradient size, enables the same accuracy as the no-compression baseline for a similar iteration count [117]. Let $b \in \mathbb{R}^d$ be a bitmask with '1' bit in $b[i]$ indicates the corresponding element $x[j,i]$ in the j^{th} input feature is selected. The element-wise product, $x \odot b$ generates a sparse vector of the original input feature vector, x. The sparse vector can be represented as (*index, value*) pairs—one contains the value of the selected element of x, and the other contains the indices where $b[i] = 1$. Based on this, we can use popular sparsification approaches, such as Top-k (bitmask b is selected with $b[i] = 1$ if $|x[i]|$ is the k largest absolute value element of x) or Random-k (randomly select a set of k indices out of d and the k corresponding bits of b are set to '1') in the feature space of every row vector of X to reduce the embedding dimension and by sending only (*index, value*) pairs to perform efficient attention by reducing d multiplications to k, where $k << d$. (*b*) **CUR type matrix decomposition for input features.** A rank-k CUR approximation of an input matrix, X, is given by $X \approx CUR$, such that C is made from k columns of X, R is made from k rows of X, and $U \in \mathbb{R}^{k \times k}$. This way, we can obtain a lower dimensional feature embedding while still keeping some input feature vectors from the original input X. (*c*) **Nystorm approximation of input features** can be another fruitful strategy. (*d*) **Sketch and Project.** We can use a random sketch of the original matrix, X by drawing a random matrix S such that $S = I_{:R}$ be a column concatenation of the columns of the $n \times n$ identity matrix I indexed by R and obtain $X_{R:}$; or $S = I_{:C}$, a column concatenation of the columns of I indexed by C, which is a random subset of $[d]$ and obtain $X_{:C}$. (*e*) **Low-rank+sparse) decomposition [51, 52]** is unexplored in the encoder-decoder architecture and our initial theoretical investigation indicates this has a lot of potential. Note that, the value, V, query, Q, and key, K matrices for the transformer's encoder structure are obtained via the linear transformations of X via random matrices, W_V, W_Q, and W_K, respectively. Hence, according to the famous Johnson–Lindenstrauss Lemma [104], they are inherently low-rank. Nevertheless, this part is highly ignored in the ML community and we plan to explore in this direction.

Initial results. PI Dutta has initiated the preliminary investigations on task (*a*) and (*e*) explained above. While a baseline ViT [46] with a batch size of 16 achieves 83.48% test accuracy on the CIFAR-10 dataset, removing 40% of the features of K and Q in different iterations achieves 83.25% test accuracy, while removing an 80% of them achieves 79.86% test accuracy. On the other hand, by reducing the size of the support set of $QK^\top$ to 50% the test accuracy drops only to 82.51%, and the model maintains an 80.425% test accuracy while reducing the size of the support set to 90%. We reckon these are remarkable results and cannot present all variations of masking strategies during the training due to space limitations.

5 Evaluation

For initial studies, the PIs will conduct preliminary evaluations of the proposed methods on the existing drone dataset, referred to as MAVREC 1.0. For a comparison to the state-of-the-art, our single-view semi-supervised approach will be tested on public datasets, including visdrone [196], DOTA [179], AU-AIR [18], and MOHR [190]. To ascertain the broader applicability of our view-generation viewpoint knowledge transfer methodology, we will extend our evaluation to downstream tasks using datasets like NTU-RGB+D [154] and Toyota Smarthome [34] for action recognition. The Pouring dataset [153] will be leveraged for evaluating human imitation learning as a downstream task.

For the aerial video conversational framework, evaluations will use MAVREC, and the resultant model's generalizability will be tested on UCF ARG [134] for zero-shot object tracking and recognition. To augment the MAVREC 1.0 dataset's diversity, public datasets across varied geographies will be integrated, allowing geographic awareness assessments of the outcome of the hierarchical training of local aerial visual models. To evaluate the efficacy, efficiency, and generalizability of our geography-aware aerial foundation model, we will test it on anomalous aerial videos from the web, encompassing scenarios such as street crimes, wildfires, and war zones. This evaluation should underscore the significance of our proposed research and highlight its broad applicability. For efficient training and to evaluate personalization and modality collaboration in

our FL settings, we plan to go beyond the multi-GPU data center training. We plan to access geographically remote cloud computing servers/platforms to deploy our multimodal, efficient FL training, and realize the potential of the in-situ hierarchical training paradigm across varying network bandwidths.

6 Broader Impacts

Scientific impact. This proposal presents real-world challenges and novel approaches for visual representation learning of images captured from aerial perspective. Advances resulting from this project will provide new benchmark dataset, models and algorithms for better representation of images and videos from drone videos. The advances are likely to benefit the computer vision community towards developing foundation models for domains constrained by limited data and geographic dependencies. Beyond its immediate implications for the research community, the proposed research holds promise for significant impact in the domain of AI for Climate, as evidenced by its potential for early wildfire detection, with simulations demonstrating a near 99% accuracy [19]. In addition, their application can overcome barriers and accelerate broader use of drones in agriculture and natural resources. Furthermore, the application of such techniques has the capacity to surmount obstacles and expedite the broader adoption of drones in agricultural and natural resource management [79].

Student research training and support. All the PIs are committed to integrate the research activities into an educational plan to train future researchers and practitioners. This project will provide training and support for two graduate students and one undergraduate student (we will request REU funding), who will collaborate and receive mentoring from the PIs. Support for this project will allow the students to attend the primary vision and machine learning conferences to present the research findings, develop strong communication skills, and get helpful feedback from the community.

Teaching and curriculum development. The project entails complementary educational activities aimed at engaging undergraduate and graduate students in research. Based on the research findings, the PIs will develop and teach a new lecture on"*object detection in aerial perspective*" for both undergraduate and graduate computer vision courses at their respective universities. Course materials will be publicly shared online on each PI's course webpages. The PIs have organized several workshops and special sessions. Additionally, they plan to submit a workshop proposal on aerial visual understanding and host a challenge on MAVREC dataset for CVPR 2025 to benefit the research community. The PIs will further increase their contribution to the community with such synergistic activities.

Outreach activities. PIs will participate in outreach educational activities to engage K-14 students in science, engineering, and math. PI Das will serve as a mentor to students in the REU summer research program, which attracts, supports, and develops the abilities and potential of academically talented and outstanding individuals from underrepresented groups. PI Dutta will involve female undergraduate through Summer Undergraduate Research Fellowship and independent studies. PI Dutta will work with the Office of Diversity Education and Training at UCF to identify and recruit women for research experiences through different research programs.

7 Prior NSF Support

PI SDas recently started receiving support under NSF IIS-2245652 grant: "CRII: RI: Understanding Activities of Daily Living in Indoor Scenarios" (Amount: $175,000, Period: 8/01/2023-7/31/2025). *Intellectual Merit:* This ongoing project seeks to develop a multi-modal framework for recognizing Activities of Daily Living (ADL) for monitoring elderlies. This project addresses the challenge of scarcity of available video data in the ADL domain and how that can be solved through self-supervised and multi-modal representation learning. *Broader Impacts:* This project performs complementary educational educational and outreach activities that engage students in research and STEM. It has begun to train graduate, undergraduate, and high-school students.
PI Dutta is a starting investigator and does not have any prior NSF support.