Mapping Cultural Ecosystem Service Flows from Social Media Imagery with Vision–Language Models: A Zero-Shot CLIP Framework

Hao-Yu Liao[a], Chang Zhao[a, *], Caglar Koylu[b], Haojie Cao[c], Jiangxiao Qiu[d], Corey T. Callaghan[e], Jiayi Song[c], Wei Shao[f, g]

[a] Agronomy Department, University of Florida, Gainesville, Florida, United States

[b] School of Earth, Environment, and Sustainability, University of Iowa, Iowa City, United States

[c] School of Natural Resources and Environment, University of Florida, Gainesville, Florida, United States

[d] School of Forest, Fisheries, and Geomatics Sciences, Fort Lauderdale Research and Education Center, University of Florida, Gainesville, Florida, United States

[e] Department of Wildlife Ecology and Conservation, Fort Lauderdale Research and Education Center, Institute of Food and Agricultural Sciences, University of Florida, Gainesville, Florida, United States

[f] Department of Medicine, University of Florida, Gainesville, Florida, United States

[g] Department of Electrical and Computer Engineering, University of Florida, Gainesville, Florida, United States

[*]Corresponding Author: Chang Zhao (email: changzhao@ufl.edu)

Abstract

Geotagged social media imagery provides a valuable source for mapping cultural ecosystem service (CES) flows, which represent realized human interactions with nature, yet its open-world user-generated content poses challenges to automated content analysis. Supervised models require large labeled datasets and show limited generalization across contexts, whereas unsupervised approaches often need post-hoc interpretation. Vision–language models offer a promising alternative but remain largely unexplored in CES research. We present a label-efficient framework that leverages the open-source Contrastive Language–Image Pretraining (CLIP) model to classify and map 12 CES flows across Florida using only 120 labeled images. Five CLIP variants and three prompt strategies were benchmarked to evaluate zero-shot performance under closed-set conditions, and three CLIP-based pipelines with differing supervision levels were compared to address the open-set challenge of filtering irrelevant content. Mixed class-specific prompts increased closed-set accuracy to 97%. Under open-set conditions, a hybrid pipeline combining a lightweight binary classifier with zero-shot CLIP inference achieved the strongest performance (accuracy = 88%; F1-macro = 0.88; F1-other = 0.91), demonstrating major gains in label-efficiency and open-set robustness. Statewide flow maps reveal consistent hotspots for outdoor recreation, wildlife viewing, and landscape aesthetics along coastal areas and major inland greenspaces, extending beyond formal park systems into urban greenspaces and other natural and working lands. The resulting map products and interactive web application provide actionable tools for identifying CES hotspots and the landscapes that support human–nature interactions.

Overall, this study demonstrates the transformative potential of foundation VLMs for large-scale CES assessment using social media imagery.

Keywords: cultural ecosystem services, CLIP, vision–language model, zero-shot learning, open-set recognition, social media imagery, natural and working landscapes.

1. Introduction

Following the Millennium Ecosystem Services Assessment (MA, 2005), the quantification and mapping of ecosystem services have gained growing attention across both scientific and policy communities worldwide over the past two decades (Naidoo et al., 2008; Maes et al., 2012; Burkhard and Maes, 2017; Edens et al., 2022). Among the diverse categories of ecosystem services, cultural ecosystem services (CES) represent the intangible, non-material benefits people obtain from ecosystems through spiritual enrichment, cognitive development, reflection, recreation, and aesthetic experience (MA, 2005). As ecosystem services connect ecological processes and functions to human benefits, the term 'ecosystem service flows' has been conceptualized in various ways (Wang et al., 2022; Li and Wang, 2023). It has been used to describe the spatial transfer or transmission of services from ecosystems to people (Bagstad et al., 2013), as well as the spatial relationships between service provisioning areas and service benefiting areas (Fisher et al., 2009; Serna-Chavez et al., 2014). In this study, we adopt a definition that distinguishes between potential supply and actual use, conceptualizing flows as the realized or *de facto* use of services by beneficiaries (Villamagna et al., 2013; Schröter et al., 2014; Burkhard et al., 2014; Baró et al., 2016; Langemeyer et al., 2018; Karasov et al., 2022).

Spatially explicit assessments of CES flows, such as identifying where and to what extent people engage in wildlife viewing (Koylu et al., 2019) or seek landscape aesthetics (Langemeyer et al., 2018), have become an increasingly important component of integrated ecosystem services frameworks. These approaches are indispensable for incorporating socio-cultural values and benefits into ecosystem accounting (Edens et al., 2022; Vallecillo et al., 2019), conservation planning, and landscape management (Plieninger et al., 2015). Compared to provisioning and regulating services, CES remain particularly challenging to quantify due to their subjective, non-material nature, which is shaped by complex socioecological relationships and dynamic human perceptions and behaviors (Huynh et al., 2022). As a result, CES are highly context-dependent and often lack consistent, systematic empirical observations across broad spatial and temporal scales (Kosanic and Petzold, 2020).

A wide range of transdisciplinary methods has been used to assess CES flows, drawing on methods and data from social science, ecology, and data science (Daniel et al., 2012; Plieninger et al., 2013; Paracchini et al., 2014; Scowen et al., 2021). At finer scales, non-monetary stated preference methods, such as questionnaires, in-depth interviews, focus groups, and participatory mapping, are most frequently used to assess CES (Cheng et al., 2019). In recent years, revealed preference methods based on passively crowdsourced social media data have emerged as a novel and increasingly prominent means of assessing CES across large spatial scales (Ghermandi and Sinclair, 2019; Havinga et al., 2020). Users upload georeferenced images, texts, and videos to platforms such as Flickr, Instagram, and iNaturalist, providing otherwise unavailable insights into where, when, and how people interact with nature and experience CES.

Geotagged Flickr images have been widely used and validated as a reliable proxy for surveyed visitation rates at various natural and cultural sites (Wood et al., 2013; Keeler et al., 2015; Sonter et al., 2016; Ghermandi, 2022). Average annual visitation estimates derived from Flickr show strong correlations with temporal patterns of on-site recreational use at national parks in the U.S. and Germany (Sessions et al., 2016; Sinclair et al., 2020). Two indicators are commonly used to quantify CES flows from Flickr data: the number of images (Willemen et al., 2015) and photo-user-days, defined as the number of unique users who took at least one photograph at a given location on a given day (Wood et al., 2013). Earlier studies typically relied on manual image

content analysis to interpret image features and identify specific CES flows across multiple spatial scales (Richards and Friess, 2015; Martínez Pastur et al., 2016; Oteros-Rozas et al., 2018).

Recent advancements in machine learning, particularly deep learning, have enabled scalable automated content analysis for CES flows (Langemeyer et al., 2023). Supervised approaches typically frame CES detection as an image classification task, fine-tuning convolutional neural networks (CNNs) such as VGG16 (Simonyan and Zisserman, 2015) and ResNet (He et al., 2016) pretrained on ImageNet (Deng et al., 2009) or Places365 (Zhou et al., 2018). These models have been applied to classify Flickr images into CES categories (Havinga et al., 2021a; Cardoso et al., 2022; Lingua et al., 2022; Winder et al., 2022; Havinga et al., 2023; Chen et al., 2025; You et al., 2025; Comalada et al., 2025; Xu et al., 2025). While demonstrating strong performance, CNN-based models typically require large volumes of manually labeled training data for effective transfer learning (Torrey and Shavlik, 2010), which limits scalability, especially for rare CES classes. In addition, these models are inherently constrained by the predefined categories and the geographic or cultural contexts represented in the training data, reducing their adaptability to novel CES classes or new spatial settings.

To reduce reliance on labeled data, a number of studies have explored unsupervised approaches by applying clustering algorithms, such as K-means (Huai et al., 2022) and hierarchical clustering (Richards and Tunçer, 2018; Lee et al., 2019; Goldspiel et al., 2023; Yee and Carrasco, 2024), as well as topic modeling techniques (Egarter Vigl et al., 2021) to machine generated keywords and tags produced by commercial cloud vision services such as Google Cloud Vision, Microsoft Azure Cognitive Services, and Clarifai. These approaches can reveal latent patterns and offer greater flexibility for CES classification across diverse contexts. However, the machine generated tags typically require human interpretation to translate objects or scene attributes into meaningful CES themes (Cao et al., 2022; Richards and Lavorel, 2022; Chai-allah et al., 2025), and the resulting clusters often do not align well with standardized CES typologies such as the Common International Classification of Ecosystem Services (CICES) (Haines-Young, 2023). Moreover, commercial tagging services show substantial variation in how they identify and characterize CES-related content (Ghermandi et al., 2022), and their full-featured versions are not freely accessible, limiting the scalability and reproducibility of large-scale CES assessments.

In contrast to the aforementioned learning paradigms, foundation vision–language models (VLMs) trained on paired image–text data have recently transformed computer vision (Bordes et al., 2024; Zhang et al., 2024) and offer strong potential for transferable, label-efficient analysis of social media image content in CES assessments. By aligning visual and textual representations in a shared semantic space, VLMs support zero-shot classification, enabling recognition of categories without task-specific retraining. A leading example is the open-source Contrastive Language–Image Pre-training (CLIP) model (Radford et al., 2021). Trained on hundreds of millions of web-scale image–text pairs through contrastive learning, CLIP embeds images and natural-language prompts into a shared semantic space and infers the most relevant category based on cosine similarity between their embeddings. This multimodal design enables CLIP to associate images with CES-related categories expressed as natural-language prompts (e.g., “This is an image of wildlife viewing”), making prompt-based CES classification both scalable and transferable across diverse contexts. Recent environmental applications of CLIP, including assessing perceptions of park accessibility and quality from social media (Zhao et al., 2024) and identifying species from iNaturalist data (Liu et al., 2024), highlight its potential for CES research.

CLIP's open-vocabulary capability provides a strong foundation for CES classification, though its performance depends on how textual prompts are formulated. Prompt engineering, the design of natural-language descriptions to guide model interpretation, has become central to the effective use of VLMs (Zhou et al., 2022a). Prior studies show that well-crafted prompts can enhance zero-shot classification accuracy across various domains (Yong et al., 2023; Shtedritski et al., 2023; Levering et al., 2024; Russe et al., 2024). In CES-related work, Luo et al. (2025) evaluated prompting strategies using large language models (LLMs) to classify CES from text reviews on the Ctrip platform. They compared task-defining, CES-descriptive, location-aware, and keyword-assisted prompts, and found that engineered prompts outperformed non-prompted baselines for several CES categories, underscoring the importance of class-specific prompt design for CES applications. Yet, the potential of VLMs such as CLIP for image-based CES classification remains largely unexplored. Given that CLIP infers labels by comparing images with CES-related textual descriptions, it is important to examine how prompt specificity, from concise to detailed or mixed descriptions, influences CES identification.

Beyond prompt sensitivity, CLIP exhibits limitations when applied to open-world social media imagery, where many photos are unrelated to CES or represent CES beyond the targeted categories. Although CLIP appears open-vocabulary, it still operates under a closed-set assumption by comparing each image to a finite, prompt-defined set of CES classes, often forcing irrelevant content into one of the known categories (Miller et al., 2025). Under closed-set conditions, the model cannot reject out-of-scope imagery, leading to systematic misclassification of non-CES photos (e.g., selfies, advertisements) and CES types not represented in the prompt set—an open-set recognition challenge shared by CLIP and many prompt-based VLMs. Compounding this, CLIP's limited capacity to interpret linguistic negation (Quantmeyer et al., 2024; Park et al., 2025; Kang et al., 2025) complicates the construction of effective negative prompts for representing the broad "other" class. These challenges underscore the need for explicit open-set filtering mechanisms when applying CLIP to noisy social media data.

To address these challenges, this study presents one of the first large-scale applications of CLIP with open-set filtering for CES classification and flow mapping from social media imagery. We developed a scalable and label-efficient CLIP-based framework to map 12 CES across Florida's natural and working lands (NWLs), including 10 outdoor recreation, wildlife viewing, and landscape aesthetics categories, along with an 'other' class for irrelevant content. We evaluated three text-prompt strategies with varying semantic richness to assess CLIP's zero-shot performance under closed-set conditions. We then benchmarked three modeling pipelines that differ in their use of CLIP's zero-shot capability and level of supervision to address open-set recognition: (1) a minimally supervised zero-shot model with class-specific cosine similarity thresholds; (2) a hybrid pipeline that integrates a lightweight binary classifier to filter irrelevant images while retaining zero-shot classification for CES categories; and (3) a fully supervised multi-class classifier trained on CLIP embeddings for all classes. These comparisons demonstrate how CLIP can be leveraged as a foundation VLM to achieve accurate and robust CES classification under label-efficient learning regimes. Specifically, our study addressed the following questions:

(1) How effectively can CLIP classify CES using zero-shot learning under closed-set conditions, and how does its performance vary across simple, detailed, and mixed prompt-design strategies?

(2) How do the three CLIP-based modeling pipelines differ in accuracy and open-set robustness when classifying CES from highly variable social media imagery under low-label conditions?
(3) What spatial patterns of CES flows emerge across Florida's natural and working lands, and how are these flows associated with underlying land-use/land-cover compositions?

2. Methodology

Figure 1 illustrates the overall research framework, which includes: (1) retrieval of geotagged Flickr images and spatial filtering to Florida's NWLs; (2) CES classification using three CLIP-based modeling pipelines, followed by model evaluation and comparison; and (3) spatial analysis and visualization of CES flows.

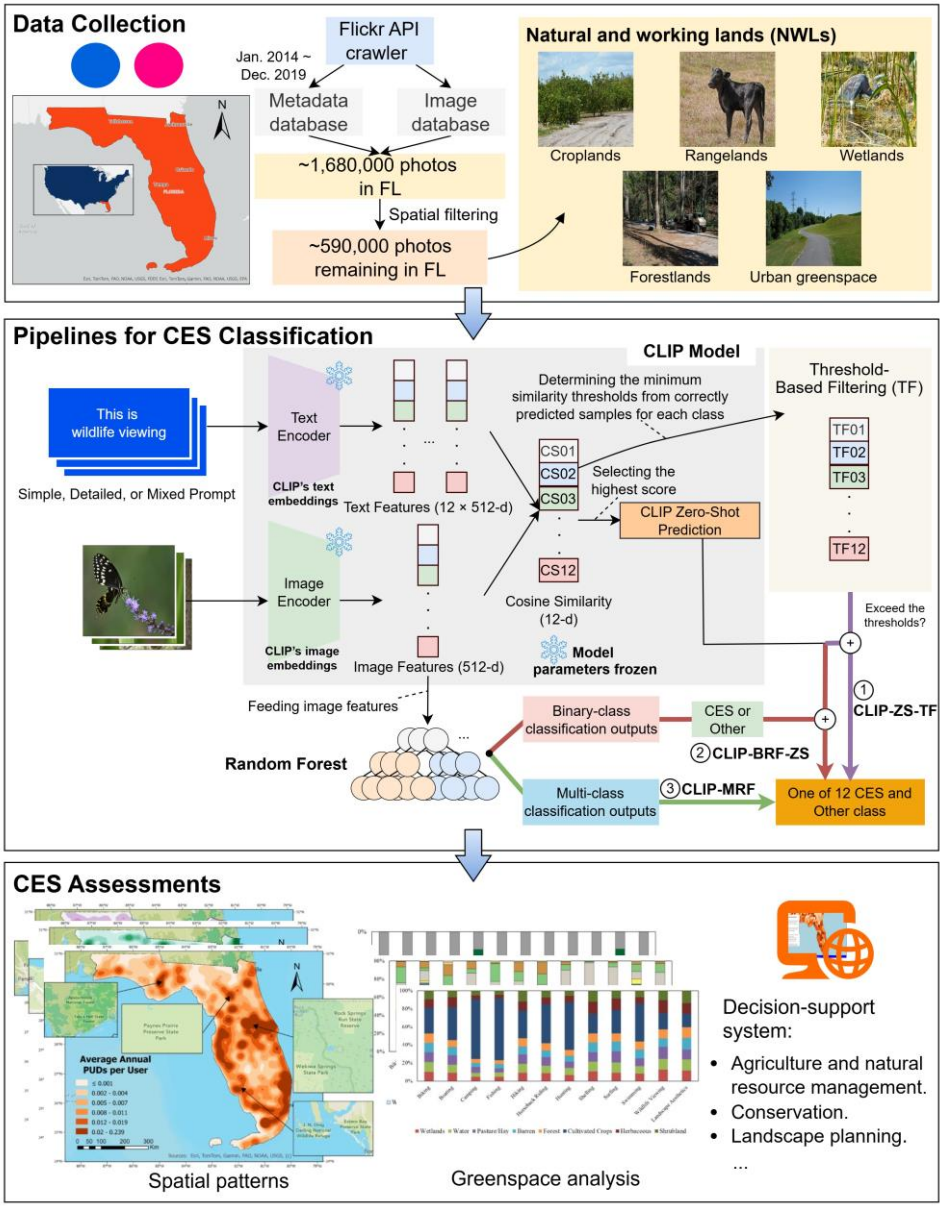

Figure 1. Research framework leveraging geotagged social media images and CLIP-based vision–language modeling for cultural ecosystem service assessment.

2.1 Flickr Data Collection and Spatial Filtering

Approximately 1.68 million of geotagged Flickr images taken between 1 January 2014 and 31 December 2019 in Florida were retrieved by querying the Flickr Application Programming Interface (API). The retrieval period ended in 2019 to avoid the influence of COVID-19-related regulations on human mobility patterns. Each image included spatiotemporal metadata essential for estimating CES flows: (1) a unique photo ID, (2) a unique owner ID identifying the user, (3) the date the photo was taken, and (4) geographic coordinates (latitude and longitude).

Spatial filtering was applied to retain only images located within Florida's NWLs, which encompass a diverse mosaic of wetlands, croplands, rangelands, forestlands and urban greenspaces (*U.S. Climate Alliance*, 2022). Specifically, images located in urban areas (U.S. Census Bureau, 2020) were excluded unless they overlapped with national parks, state parks, protected areas, or other greenspaces in urban environments. Spatial boundaries for filtering were derived from multiple sources: state park boundaries from the Florida Department of Environmental Protection (FDEP, 2025), national parks from the National Park Service (NPS Land Resources Division, 2025), protected areas from the U.S. Geological Survey's Gap Analysis Project (USGS GAP, 2024), and urban greenspaces from the ParkServe database (Trust for Public Land, 2025). The filtering process resulted in about 590,000 images for CES analysis.

2.2 Manual Image Labeling

We focused on three broad CES categories: (1) nature-based outdoor recreation, (2) wildlife viewing, and (3) landscape aesthetics. Within the recreation category, we identified 10 most common recreational activities in Florida based on state agency reports from the Florida State Parks (FDEP, 2025). A total of 4305 Flickr images were manually labeled across these 12 CES classes, along with an additional "other" class representing primarily irrelevant or noisy content (e.g., indoor or non-nature-related scenes), though it may also include CES types not explicitly addressed in this study (e.g., cultural heritage or spiritual inspiration). The labeled dataset was divided into two subsets: the development set (n = 120) and the testing set (n = 4185), as shown in Table 1. Each CES class contained at least 180 labeled images, with only 5 deliberately allocated for prompt tuning and supervised learning to represent low-label conditions in CES classification.

Table 1. Number of labelled geotagged Flickr images in the study.

	Broad CES Category	CES Classes	Development Set	Testing Set
1	Outdoor Recreation	Biking	5	182
2		Boating	5	312
3		Camping	5	234
4		Fishing	5	251
5		Hiking	5	478
6		Horseback Riding	5	248
7		Hunting	5	177
8		Shelling	5	183
9		Surfing	5	191
10		Swimming	5	425
11	Wildlife Viewing	Wildlife Viewing	5	380
12	Landscape Aesthetics	Landscape Aesthetics	5	404
13	Other*		60	720
	Total		120	4185

* The "other" class includes images that do not belong to any of the 12 targeted CES classes.

2.3 CLIP Model

The open-source CLIP model enables effective zero-shot classification by jointly aligning visual and textual representations within a shared embedding space (Radford et al., 2021). It consists of a visual encoder and a text encoder trained jointly via contrastive learning, which maximizes the cosine similarity between matched image–text pairs while minimizing it for mismatched pairs. Formally, given an image I and text description T:

$$v = f_\theta(I) \tag{1}$$

$$u = g_\theta(T) \tag{2}$$

$$v, u \in R^d \tag{3}$$

where $f_\theta(\cdot)$ and $g_\theta(\cdot)$ denote the image and text encoders, respectively, and v and u are their corresponding embedding vectors. To normalize the embeddings, the L2-norm is applied:

$$\hat{v} = \frac{v}{\|v\|} \tag{4}$$

$$\hat{u} = \frac{u}{\|u\|} \tag{5}$$

The cosine similarity between the image and text embeddings is then computed as:

$$CosSim(I, T) = \hat{v}^T \hat{u} = \frac{v^T u}{\|v\|\|u\|} \tag{6}$$

Given an image I with k candidate text prompts, the predicted class is obtained by selecting the prompt with the highest cosine similarity:

$$C_I = \underset{k}{\operatorname{argmax}} \frac{v^T u_k}{\|v\|\|u_k\|} \tag{7}$$

where C_I denotes the predicted class for image I.

CLIP has demonstrated excellent zero-shot performance on standard benchmarks such as ImageNet-1k and Caltech-101 without requiring task-specific retraining (Cherti et al., 2023; Ilharco et al., 2025). For this study, we used MobileCLIP-S1 (Vasu et al., 2024), a state-of-the-art CLIP variant that integrates a hybrid CNN-Transformer architecture optimized for efficient inference and high classification accuracy. To benchmark performance across architectures and model scales, we additionally evaluated four representative CLIP variants using the same natural language prompts on the development set: ViT-B/32 (Dosovitskiy, 2020; Radford et al., 2021) and RN101 (He et al., 2016; Radford et al., 2021), EVA-01-CLIP-g/14 (Sun et al., 2023), and convnext_base_w (Liu et al., 2022; Tan et al., 2025). The classification process involved computing cosine similarity between each test image and a predefined set of simple text prompts (Table 2), assigning the label corresponding to the highest similarity score. All models were implemented in PyTorch (Paszke et al., 2019) and executed on the University of Florida's HiPerGator supercomputing cluster with a Blackwell B200 GPU, eight Intel Xeon Platinum 8570 56-core processors, and 64 GB of RAM.

2.4 CES Prompt Tuning

We developed three types of text prompts: simple, detailed, and mixed, for each of the 12 CES classes (Table 2). Simple prompts consisted of short, generic labels (e.g., "*This is hiking*"), whereas detailed prompts provided richer semantic descriptions capturing the visual and contextual characteristics of each CES class. Mixed prompts combined elements of both, retaining the simple form or adopting the detailed version depending on which yielded better class-specific performance on the development set. To ensure conceptual alignment between text descriptions

and visual representations, the detailed prompts were informed by visual patterns observed in the development set, conceptual definitions from the Common International Classification of Ecosystem Services (CICES v5.2; Haines-Young, 2023), and insights from prior CES literature (Rosenberger et al., 2017; Mitten et al., 2018; Lingua et al., 2022, 2023; Winder et al., 2022). These prompt variants allowed us to examine how varying levels of semantic richness influence CLIP's zero-shot performance in classifying CES-related images.

Table 2. Simple, detailed, and mixed text prompts used for CES classification with CLIP across the 12 CES classes.

	CES Class	Simple prompt	Detailed Prompt	Mixed Prompt
1	Biking	This is biking.	This is an image of a person riding a bicycle or mountain bike off-road and over rough or scenic terrain.	This is biking.
2	Boating	This is boating.	This is an image of boating with powerboats or human-powered vessels such as rowboats or paddle boats.	This is an image of boating with powerboats or human-powered vessels such as rowboats or paddle boats.
3	Camping	This is camping.	This is an image of a person or a tent set up in a campground, camping vehicle, temporary shelter, or cabin.	This is camping.
4	Fishing	This is fishing.	This is an image of people near a river, lake, pond, canal, wetland, or reservoir, fishing with gear such as a rod, reel, and bait.	This is fishing.
5	Hiking	This is hiking.	This is an image of people hiking or taking a casual walk on challenging scenic trails or through breathtaking countryside.	This is an image of people hiking or taking a casual walk on challenging scenic trails or through breathtaking countryside.
6	Horseback Riding	This is horseback riding.	This is an image of a person riding a horse for recreation or standing near a horse trailer.	This is horseback riding.
7	Hunting	This is hunting	This is an image of a hunter posing with a large kill, cleaning or gutting the carcass of a feral animal.	This is an image of a hunter posing with a large kill, cleaning or gutting the carcass of a feral animal.
8	Shelling	This is shelling.	This is an image of a person collecting seashells or seashells displayed on the beach.	This is shelling.
9	Surfing	This is surfing.	This is an image of people surfing on ocean waves or holding a surfboard on the beach.	This is an image of people surfing on ocean

				waves or holding a surfboard on the beach.
10	Swimming	This is swimming	This is an image of people wearing swimsuits while swimming in a pool, lake, or ocean.	This is an image of people wearing swimsuits while swimming in a pool, lake, or ocean.
11	Wildlife Viewing	This is wildlife viewing.	This is an image without people, showing wildlife such as birds, flowers, or fungi in a natural environment.	This is an image without people, showing wildlife such as birds, flowers, or fungi in a natural environment.
12	Landscape Aesthetics	This is landscape aesthetics.	This is an image without people, showing a natural, soothing, and vibrant panoramic landscape.	This is an image without people, showing a natural, soothing, and vibrant panoramic landscape.

2.5 CLIP-based Modeling Pipelines under Open-Set Conditions

Open-ended social media imagery contains substantial content unrelated to the targeted CES categories. Robust CES classifiers must therefore not only recognize relevant classes accurately but also effectively reject inputs outside the predefined query set. To assess classification accuracy and robustness, we compared three CLIP-based pipelines that share a common embedding backbone but differ in their supervision level and classification design. The first two pipelines employ CLIP's zero-shot image–text alignment for CES classification, whereas the third uses CLIP solely as a fixed image feature extractor for supervised learning with a few-shot setting. All pipelines used the same small, labeled development dataset (n = 120) for threshold or classifier calibration and were evaluated on 4,185 labeled test images (Table 1), ensuring consistent and comparable performance assessment.

2.5.1 Zero-Shot with Threshold Filtering (CLIP-ZS-TF)

This pipeline leverages CLIP's zero-shot mechanism to assign each image to one of 12 CES categories based on the cosine similarity between image and text embeddings generated from the optimized prompts. To handle open-set conditions, we applied a predictive-uncertainty filter using class-specific cosine similarity thresholds. These thresholds were empirically derived from the labeled development set by analyzing the similarity distribution of true positives within each class and selecting the minimum observed similarity as the cutoff (Table S.1). During inference, CLIP predicts the most similar CES class from the query set; if the corresponding similarity is below that class's threshold, the image is classified as "other." This uncertainty-aware filtering excludes low-confidence predictions while maintaining the label efficiency and scalability of zero-shot classification.

2.5.2 Supervised Filtering with Zero-Shot Classification (CLIP-BRF-ZS)

This pipeline experiments with few-shot learning, which integrates a lightweight supervised filtering step with zero-shot CES classification. A Binary Random Forest (BRF) classifier (Breiman, 2001) was trained on CLIP's 512-dimensional image embeddings from the same development set to distinguish CES-relevant from irrelevant ("other") images. The training data

comprised 60 positive examples (5 per CES class) and 60 negative examples (Table 1). Model hyperparameters (e.g., number of trees, maximum depth) were optimized via grid search with 3-fold cross-validation (Table S.2). During inference, each image first passes through the BRF filter: images classified as "other" are excluded, while CES-relevant images are subsequently categorized into 12 CES classes using CLIP's zero-shot inference with the optimized prompt set. The BRF was implemented in Python 3.10.18 using Scikit-learn (Pedregosa et al., 2011).

2.5.3 Fully Supervised Multiclass Classification (CLIP-MRF)

As a fully supervised method, this pipeline bypasses zero-shot inference and trains a multi-class Random Forest (MRF) directly on CLIP's image embeddings from the same development dataset to predict among 13 classes (the 12 CES types plus "other"). CLIP encoders remain frozen, with supervision applied only at the classifier level. Model hyperparameters were tuned via grid search with 3-fold cross-validation (Table S.2). During inference, the MRF directly outputs the final class label, serving as a reference for the performance achievable through fully supervised learning under a few-shot setting.

2.6 Model Evaluation

Model performance was evaluated on the testing set (n = 4185) using confusion matrices and macro-level metrics: precision, recall, and F1-score (Bishop and Nasrabadi, 2006). Macro metrics average performance equally across classes, while micro metrics weight results by class size. Given class imbalance in the testing set, only macro metrics were reported to ensure a balanced evaluation across CES categories. Precision measures the proportion of correctly predicted images among all images assigned to a given class, reflecting the ability to avoid false positives. Recall measures the proportion of actual class instances correctly classified, reflecting the ability to avoid false negatives. F1-score is the harmonic mean of precision and recall, balancing both types of classification error. We also report overall accuracy, defined as the proportion of correctly classified instances across all classes. Robustness to irrelevant content was evaluated using the F1-score of the 'other' class (F1-other) and the macro-averaged F1-score across all 13 classes (F1-macro). A robust pipeline is characterized by simultaneously high F1-other and F1-macro, indicating effective rejection of irrelevant images without compromising CES classification accuracy.

2.7 CES Flows Quantification and Spatial Analysis

We applied the best-performing CLIP pipeline with the optimal model variant and prompts to classify over 590,000 geotagged Flickr images across Florida's NWLs into 12 CES classes. To reduce bias from highly active users who may upload multiple images during a single visit, user activity was summarized as photo-user-days (PUDs), which count unique users per location per day and serve as a proxy for visitation intensity (Wood et al., 2013). For each CES class c, CES flow was quantified as the average annual PUDs across all years ($\overline{PUD}_{yr}$) at a 1 km resolution:

$$\overline{PUD}_{yr}(i,c) = \frac{1}{T}\sum_{t=1}^{T} PUD_t(i,c) \tag{8}$$

where $PUD_t(i,c)$ is the total number of PUDs for class c in grid cell i during year t, and T is the number of years included in the analysis. To account for variation in user activity, we further derived per-user flow intensity ($\overline{PUD}_{user,yr}$):

$$\overline{PUD}_{user,yr}(i,c) = \frac{\overline{PUD}_{yr}(i,c)}{N_{user}} \tag{9}$$

where N_{user} denotes the total number of unique users contributing geotagged images during the study period. The resulting unit is PUD · user^{-1} · year^{-1}. Spatial patterns of both metrics were examined on a 1 km grid using kernel density estimation with a Gaussian kernel and bandwidth selected via Silverman's rule (Silverman, 1986), allowing identification of hotspots and statewide gradients. Flow volumes and per-user intensities were summarized across NWLs types (i.e., national and state parks, protected areas, urban and non-urban greenspaces, and other NWLs), and by major land use/cover (LULC) classes from the 2019 National Land Cover Database (USGS, 2024) to evaluate how different landscape compositions support total and individual CES delivery.

To facilitate exploration of these outputs, we developed an interactive online mapping platform (https://es-geoai.rc.ufl.edu/agroes-ces-clip/) that visualizes gridded CES flow surfaces and activity-specific CES bundles summarized at the Census Block Group level. The platform enables users to toggle among CES types, inspect spatial hotspots, and explore flow patterns alongside socioeconomic and demographic indicators from the U.S. Census, providing an accessible interface for researchers, planners, and stakeholders to interact with and interpret the results.

3. Results

3.1 Model Performance Across CLIP Variants

Table 3 summarizes the closed-set performance of five CLIP variants using simple prompts on the development set for zero-shot classification across the 12 CES classes (excluding the 'other' class). Among ViT-B/32, RN101, EVA-01-CLIP-g/14, convnext_base_w, and MobileCLIP-S1, the MobileCLIP-S1 model achieved the best performance across all metrics, reaching 90% overall accuracy, 0.92 precision, 0.90 recall, and 0.90 F1-score. Notably, MobileCLIP-S1 outperformed larger transformer-based models such as EVA-01-CLIP-g/14 and ViT-B/32, indicating that its hybrid CNN–Transformer architecture effectively balances semantic representation and computational efficiency. In contrast, the CNN-based RN101 and convnext_base_w models showed lower accuracies (78–80%), reflecting their reduced ability to capture CES-relevant visual–semantic relationships in zero-shot settings. Overall, MobileCLIP-S1 provided the most accurate and consistent baseline for subsequent prompt tuning and CES classification under open-set conditions.

Table 3. Closed-set zero-shot classification performance of CLIP variants on the development set across 12 CES classes using simple prompts.

Model	Accuracy (%)	Macro precision	Macro recall	Macro F1-score
ViT-B/32	87	0.88	0.87	0.86
RN101	78	0.84	0.78	0.77
EVA01-CLIP-g/14	85	0.91	0.85	0.84
convnext_base_w	80	0.84	0.80	0.80
MobileCLIP-S1	90	0.92	0.90	0.90

3.2 Effects of Prompt Types

Table 4 summarizes the closed-set zero-shot classification performance of MobileCLIP-S1 across the 12 CES classes (excluding the "other" class) using simple, detailed, and mixed prompts. Overall accuracy increased from 90% with simple prompts to 92% with detailed prompts and 97% with mixed prompts. Similar gains were observed in macro precision, recall, and F1-score, indicating that prompt design substantially affects CLIP's text–image alignment and classification accuracy. Simple prompts remained adequate for visually distinctive and context-independent

activities such as biking, camping, fishing, horseback riding, and shelling. Detailed phrasing modestly improved overall recall, suggesting that richer textual context helps CLIP recognize a broader range of semantically relevant imagery, particularly for visually or thematically overlapping categories, improving the distinguishment of swimming from other water-based recreation, landscape aesthetics from hiking, hiking from camping, and wildlife viewing from hunting. The mixed prompt strategy achieved the highest overall performance by tailoring prompt semantic richness to class characteristics, retaining concise phrasing for visually distinct activities while adding contextual detail where semantic overlap was high.

Table 4. Closed-set zero-shot classification performance of MobileCLIP-S1 on the development set across 12 CES classes (excluding the "other" class) using simple, detailed, and mixed prompts.

	Simple prompt	**Detailed Prompt**	**Mixed Prompt**
Accuracy (%)	90	92	97
Macro Precision	0.92	0.93	0.97
Macro Recall	0.90	0.92	0.97
Macro F1-Score	0.90	0.91	0.96

3.3 Model Performance under Open-Set Conditions

Among the three CLIP-based modeling pipelines, CLIP-BRF-ZS achieved the highest performance and strongest robustness to open-set conditions (Table 5), reaching 88% overall accuracy with balanced precision (0.88), recall (0.88), F1-other (0.91) and F1-macro (0.88). In comparison, CLIP-ZS-TF attained 78% accuracy (F1-other = 0.66, F1-macro = 0.80), while the fully supervised CLIP-MRF lagged behind with 69% accuracy and poor open-set robustness (F1-other = 0.55, F1-macro = 0.75).

The CLIP-MRF model exhibited unstable behavior under limited labels, with very low recall for landscape aesthetics (0.16) and frequent misclassification of irrelevant imagery (Figure 2). While CLIP-ZS-TF retained high precision for most CES classes, it suffered from poor precision for the "other" class (0.50), indicating that many valid CES images were incorrectly rejected as irrelevant content because of its conservative similarity thresholds. It also exhibited lower overall recall (0.77), particularly in fishing (recall = 0.26).

By contrast, CLIP-BRF-ZS effectively mitigated both types of errors, raising precision for "other" class to 0.86 while retaining high recall (0.96), and substantially improving recall for challenging CES classes such as fishing (from 0.26 to 0.93) and landscape aesthetics (from 0.67 to 0.85). Overall, the CLIP-BRF-ZS pipeline maintained consistently high F1-scores (0.81–0.97) across all classes, underscoring the effectiveness of integrating lightweight supervised filtering into the zero-shot CLIP framework. This approach enables reliable discrimination of CES-relevant from irrelevant imagery and robust generalization to large, open-world social media datasets under low-label conditions.

Table 5. Comparative performance of three CLIP-based modeling pipelines on the testing set under open-set conditions.

	CLIP-ZS-TF			CLIP-BRF-ZS			CLIP-MRF		
CES Class	**Precision**	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**	**Precision**	**Recall**	**F1-Score**
Biking	0.97	0.84	0.90	0.93	0.83	0.88	0.99	0.82	0.90
Boating	0.95	0.67	0.79	0.91	0.75	0.82	0.93	0.55	0.69
Camping	0.95	0.89	0.92	0.96	0.82	0.88	0.99	0.59	0.74
Fishing	0.94	0.26	0.41	0.80	0.93	0.86	0.94	0.55	0.70
Hiking	0.96	0.64	0.77	0.91	0.79	0.85	0.96	0.54	0.69
Horseback Riding	0.98	0.70	0.81	0.98	0.96	0.97	1.00	0.92	0.96
Hunting	0.85	0.81	0.83	0.80	0.83	0.81	0.94	0.64	0.76
Shelling	0.86	0.98	0.92	0.83	0.95	0.88	0.94	0.78	0.85
Surfing	0.84	0.95	0.89	0.85	0.95	0.90	0.92	0.92	0.92
Swimming	0.98	0.78	0.87	0.95	0.91	0.93	0.98	0.82	0.89
Wildlife Viewing	0.85	0.85	0.85	0.85	0.90	0.88	0.99	0.68	0.81
Landscape Aesthetics	0.91	0.67	0.77	0.86	0.85	0.85	0.73	0.16	0.26
Other class	0.50	0.97	0.66	0.86	0.96	0.91	0.38	1.00	0.55
Accuracy (%)	78			88			69		
Macro Precision	0.89			0.88			0.90		
Macro Recall	0.77			0.88			0.69		
Macro F1-score	0.80			0.88			0.75		

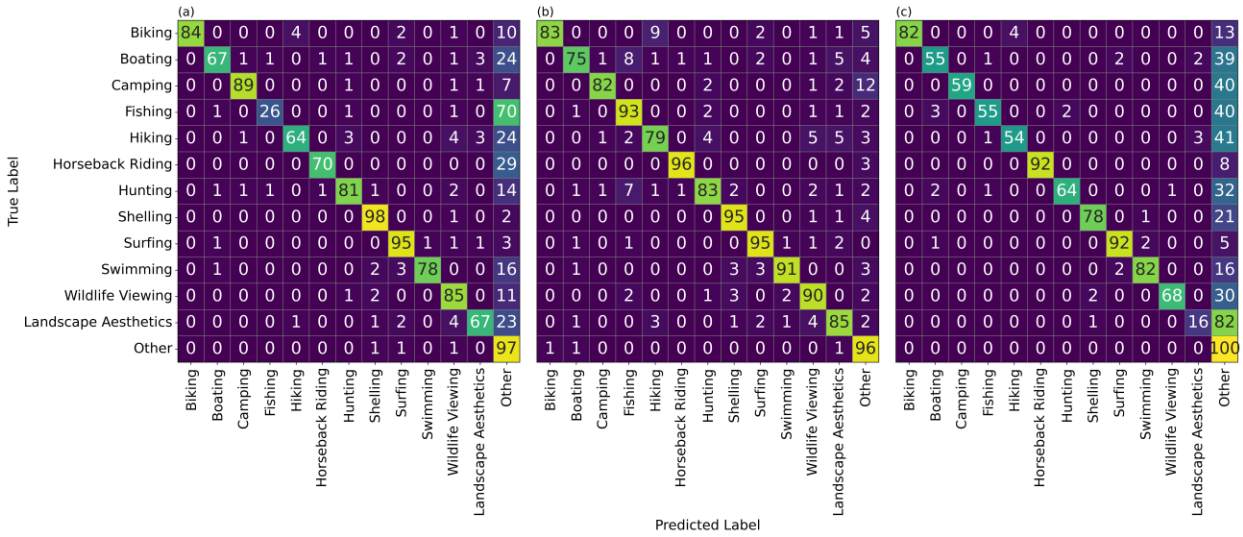

Figure 1. Normalized confusion matrices for the three CLIP modeling pipelines (a) CLIP-ZS-TF, (b) CLIP-BRF-ZS, and (c) CLIP-MRF on the testing set under open-set conditions.

3.4 CES Flow Quantification

Using the CLIP-BRF-ZS modeling pipeline with the MobileCLIP-S1 model and mixed prompts, we classified over 590,000 Flickr images into 12 CES categories across Florida's NWLs from 2014 to 2019 (Figure 3).

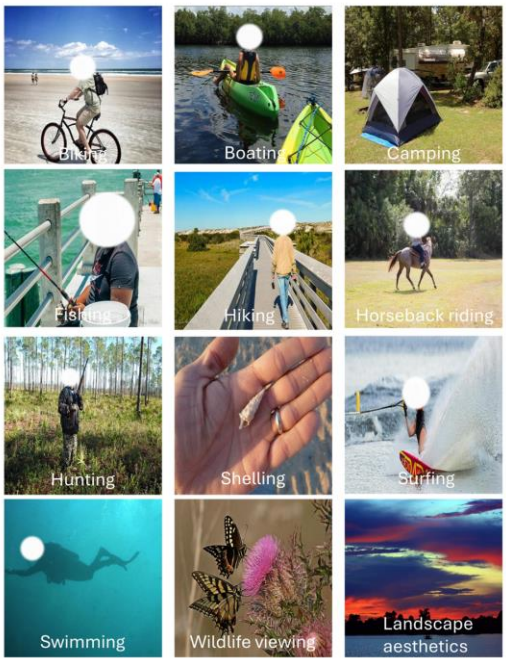

Figure 3. Example predictions for the 12 CES classes from Flickr images using the CLIP-BRF-ZS model.

Table 6 summarizes the magnitude and variability of CES flows across NWLs in Florida at 1 km resolution. Urban greenspaces exhibited the highest mean values for both flow metrics (i.e., mean $\overline{PUD}_{yr} = 1.46$, mean $\overline{PUD}_{user,yr} = 0.28$), underscoring their central role as key CES benefiting areas. State and national parks showed moderate total flow and slightly lower per-user activity intensity. Protected areas and other NWLs had lower mean totals but the highest coefficients of variation (CV > 3.6), indicating pronounced spatial heterogeneity and localized CES hotspots within these landscapes. Notably, national parks exhibited the highest per-user variability (CV = 4.67), suggesting considerable differences in visitor engagement among sites.

Table 6. Descriptive statistics of CES flows at 1 km resolution by NWL types.

NWL types	Average annual PUDs ($\overline{PUD}_{yr}$)					Average annual PUDs per user ($\overline{PUD}_{user,yr}$)				
	Min	Max	Mean	Std	CV	Min	Max	Mean	Std	CV
Urban greenspaces	0.17	107.2	1.46	4.34	2.97	0.17	13.8	0.28	0.55	1.95
Non-urban greenspaces	0.17	11.5	0.60	1.14	1.91	0.17	2.7	0.22	0.22	1.00
State parks	0.17	31.7	0.93	2.16	2.32	0.17	4.7	0.22	0.20	0.90
National parks	0.17	44.8	1.04	2.94	2.82	0.17	29.8	0.24	1.11	4.67
Protected areas	0.17	91.5	0.73	2.66	3.66	0.17	8.7	0.26	0.37	1.42
Other NWLs	0.17	75.3	0.51	1.86	3.69	0.17	32.2	0.25	0.79	3.10

Std: standard deviation, cv: coefficients of variation.

3.5 Spatial Patterns of CES Flows across NWLs

Figures 4 and 5 illustrate the spatial distribution of average annual CES flows ($\overline{PUD}_{yr}$) across Florida for outdoor recreation, wildlife viewing, and landscape aesthetics. Statewide, all three CES types exhibited broadly similar spatial gradients, with the most prominent hotspots concentrated along both coasts and around major inland greenspaces (Figure 4). This consistency indicates that many locations simultaneously delivered multiple CES, while subtle differences in hotspot intensity reflect service-specific emphases: recreation and aesthetics peak along densely visited coastal regions, whereas wildlife viewing extends more broadly across inland park systems. Figure 5a further provides localized examples showing that high-flow areas for outdoor recreation closely aligned with the distribution of national and state parks, confirming a strong spatial coupling between recreational use and protected park landscapes.

These general patterns were further differentiated when examining each broad CES category individually. For outdoor recreation (Figure 4a; Figure 5b), high flows concentrated around major coastal and aquatic destinations such as Gulf Islands National Seashore and Skyway Fishing Pier State Park, as well as near prominent inland springs and lake systems. Wildlife viewing (Figure 4b; Figure 5c) hotspots occurred primarily within protected forests and wetland mosaics, including Apalachicola National Forest, Paynes Prairie Preserve State Park, Wekiwa Springs State Park, and J.N. "Ding" Darling National Wildlife Refuge, reflecting the high biodiversity of these habitats. Landscape aesthetics (Figure 4c; Figure 5d) exhibited dense clusters across coastal, estuarine, and riverine landscapes, with notable hotspots in Fred Gannon Rocky Bayou State Park, Waccasassa Bay State Preserve, Palatka–St. Augustine State Trail, and the Everglades National Park.

Statewide, wildlife viewing (38%) and landscape aesthetics (34%) contributed the largest shares of total CES flow volumes, followed by outdoor recreation (28%). The pie charts in Figure 4 (right panels) further summarize the relative contributions of greenspace types and other NWLs to overall CES flows. In general, protected areas and urban greenspaces emerged as the dominant service benefiting areas. Urban greenspaces supported 22% of wildlife-viewing, 28% of recreational, and 27% of aesthetic flows, exceeding the individual contributions of state and national parks (10–14%). Protected areas contributed 43% of total wildlife-viewing flows, consistent with their core role in biodiversity conservation and habitat provision. Beyond designated greenspaces, CES flows extended widely across other NWLs, contributing approximately 12–22% of total statewide activity and exhibiting distinct LULC compositions. Within this group, aquatic and wetland environments dominated, together accounting for 50% of outdoor recreation, 64% of wildlife viewing, and 56% of landscape aesthetics. Barren areas, mainly beaches and coastal dunes, contributed 25% and 23% to recreation and aesthetics, respectively, emphasizing the recreational and scenic value of Florida's shoreline. Working lands also played a measurable role: cultivated crops (6–13%) were more associated with recreation, while pasture and hay fields (8–10%) contributed more evenly across all three broad CES categories, indicating that managed rural landscapes continue to provide realized cultural benefits.

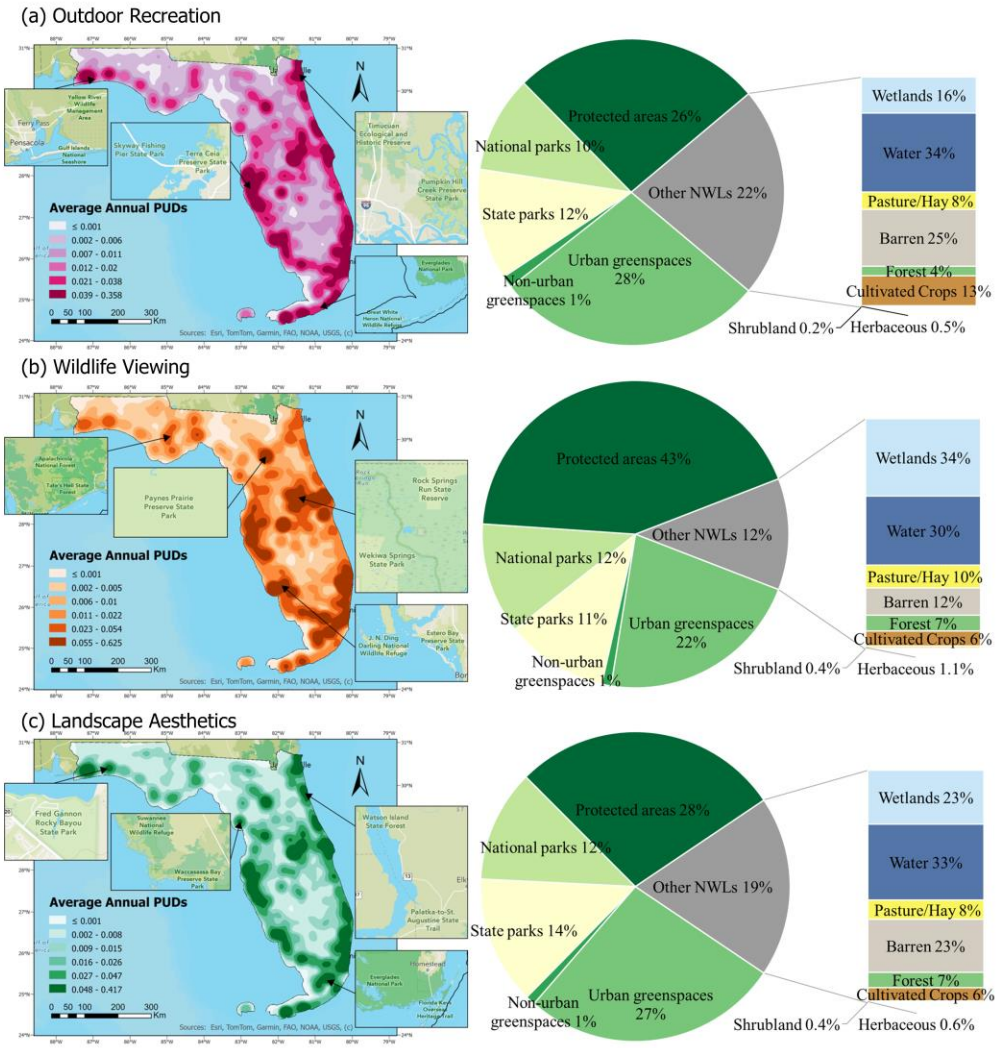

Figure 4. Kernel density estimation of average annual CES flows ($\overline{PUD}_{yr}$) at 1 km resolution across Florida for (a) outdoor recreation, (b) wildlife viewing, and (c) landscape aesthetics by NWLs.

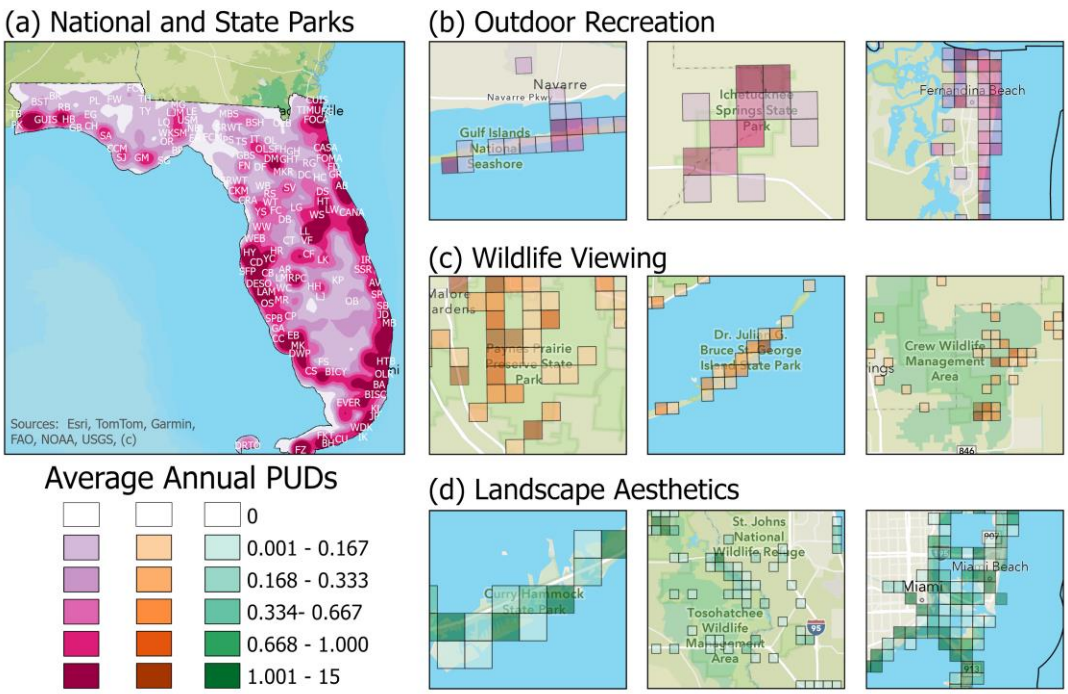

Figure 5. Spatial distribution of average annual CES flows ($\overline{PUD}_{yr}$) at 1 km resolution across Florida. (a) Kernel density of outdoor recreation flows overlaid with National and State Parks; (b–d) zoom-in views of flows for outdoor recreation, wildlife viewing, and landscape aesthetics. Park abbreviations follow FDEP (2025) and NPS Land Resources Division (2025) definitions.

Figure 6 disaggregates the broad recreation category to reveal activity-specific CES flow patterns and their distinct ecological and spatial contexts. The distribution of activities (Figure 6a) shows that other NWLs, protected areas, and urban greenspaces supported most recreational flows, while state and national parks contributed smaller but well-balanced shares across all activities. The underlying LULC composition (Figure 6b) further highlights the landscape diversity underpinning these activities. Shelling, surfing, and swimming were concentrated in barren areas corresponding to Florida's coastal beaches. Boating and fishing were closely associated with open water and wetlands, aligning with their aquatic settings. Wetlands also supported a wide range of non-aquatic activities, in particular hiking and hunting, underscoring their dual ecological and recreational importance. Working lands contributed smaller yet notable shares of recreational flows: cultivated crops were linked to general outdoor recreation, while pasture and hay fields were primarily associated with horseback riding, reflecting the combined agricultural and leisure use of these landscapes.

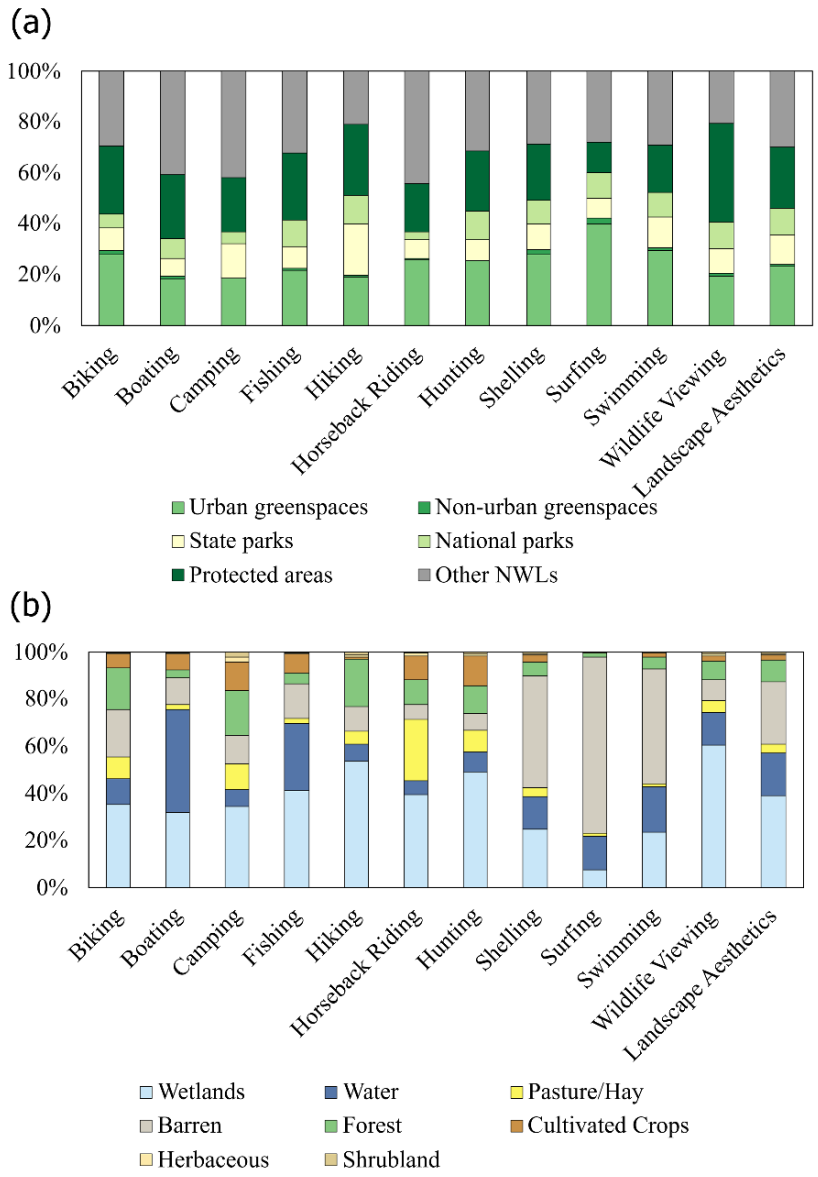

Figure 6. Activity-specific average annual CES flows ($\overline{PUD}_{yr}$) by (a) NWLs and (b) LULC types.

3.6 Per-User CES Flow Patterns

Per-user average annual CES flow intensity ($\overline{PUD}_{user,yr}$) was more evenly distributed across Florida than total flows, with moderate but consistent hotspots along both coasts and around major inland greenspaces (Figure S.1). Across NWLs, urban greenspaces exhibited the highest average per-user intensity (0.28), followed by protected areas, other NWLs and national parks (0.24–0.26), while both non-urban greenspaces and state parks showed slightly lower values (0.22; Figure 7 and Table S.3). These results indicate that intensive individual CES flow was not confined to formal park systems; rather, accessible spaces and multifunctional working landscapes often support comparable or stronger per-user interactions. Per-user intensities varied markedly across LULC types (Figure 8; Table S.4). Cultivated crops showed the highest mean per-user intensity

(0.32), driven largely by camping and fishing, highlighting the combined agricultural and recreational functions of working landscapes. Wetlands, water, and barren coastal areas exhibited moderate but consistent intensities (0.23), followed by pasture/hay (0.22), forests and herbaceous areas (0.21), and shrublands (0.18). Overall, most LULC types support a relatively diverse and balanced mix of CES flows, whereas croplands display a highly concentrated and uneven activity profile despite their high mean intensity.

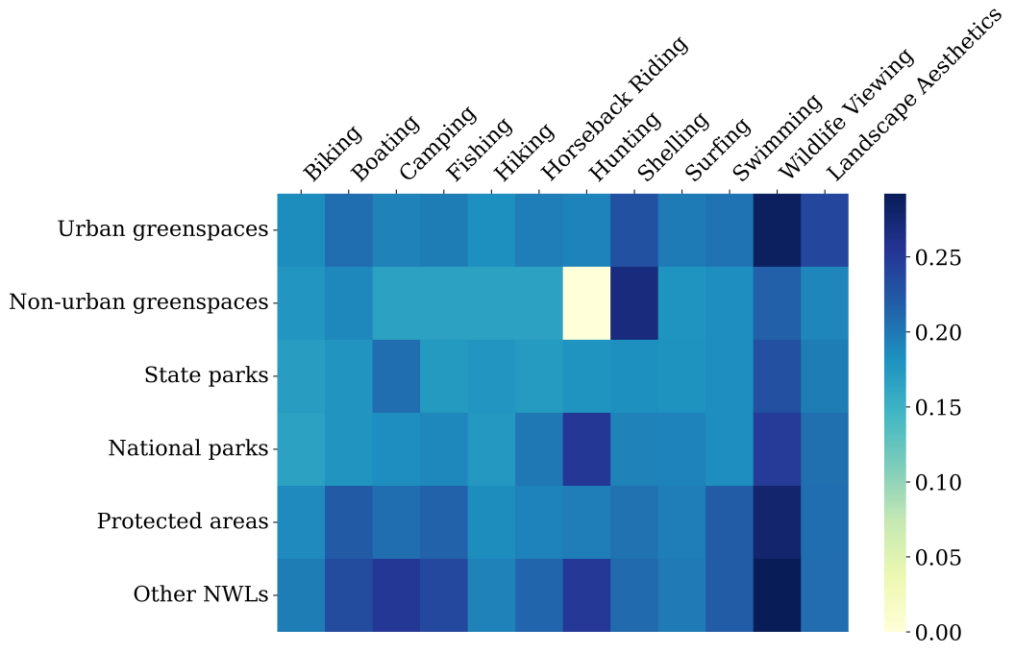

Figure 7. Per-user CES flow intensity across NWLs and CES classes. Values represent mean $\overline{PUD}_{user,yr}$.

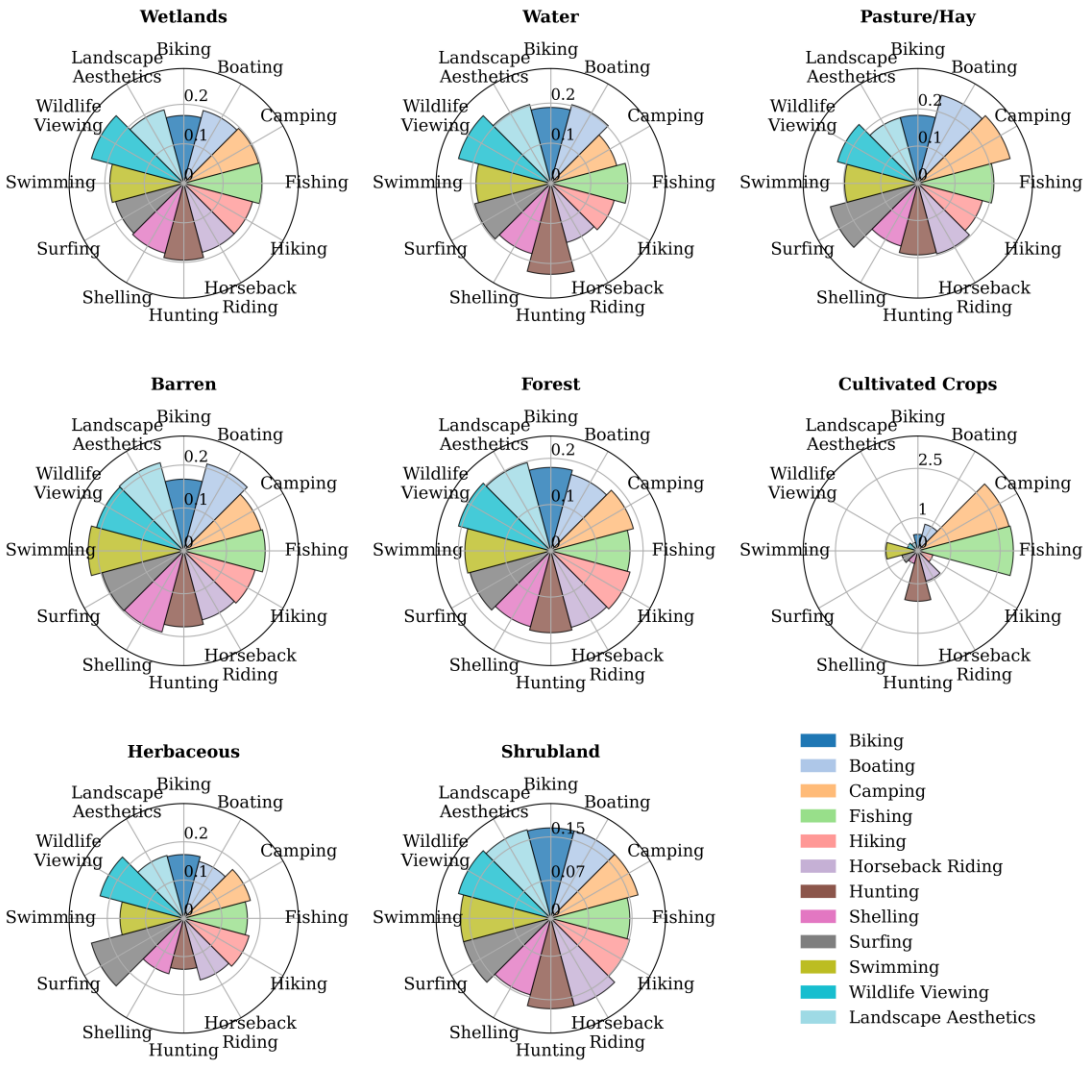

Figure 8. Per-user CES flow intensity by CES classes and LULC types. Values represent mean $\overline{PUD}_{user,yr}$.

4. Discussion

4.1 Advancing CES Classification with Vision–Language Models

Recent advances in deep learning have opened new opportunities for CES research, enabling automated analysis of large volumes of social media imagery (Ghermandi and Sinclair, 2019; Manley et al., 2022; Schirpke et al., 2023; Ghermandi et al., 2023). However, CNNs fine-tuned on custom datasets are often constrained by their reliance on large labeled datasets, fixed categorical schemes, and limited semantic interpretability, making scalability and cross-context generalization difficult, especially when applied to novel geographical settings (Cardoso et al., 2022; Winder et al., 2022). VLMs such as CLIP directly address these limitations by aligning images and text in a shared embedding space, enabling zero-shot classification and open-vocabulary recognition through natural-language prompts (Radford et al., 2021). Building on this emerging modeling paradigm, our study introduces CLIP-based frameworks that deliver label-efficient, interpretable, and transferable CES assessments across Florida's NWLs.

A key conceptual advance of our approach is the shift from closed-set, label-dependent CES classification toward language-grounded, open-set recognition. Traditional vision-only models can only predict categories on which they were explicitly trained and require substantial supervised data to add new classes (Torrey and Shavlik, 2010). In contrast, CLIP enables zero-shot classification based on natural-language descriptions of concepts, allowing categories to be flexibly defined, expanded, or adapted to new regions by modifying text prompts. This capacity for open-vocabulary recognition represents an important step forward for CES research, where categories are diverse and often difficult to pre-enumerate. Our framework also improves on unsupervised and tag-based CES assessments that relied on machine-generated labels and required post-hoc interpretation to map tags to CES themes (Egarter Vigl et al., 2021; Costadone and Balzan, 2023). Because such tags may be imprecise or inconsistent across platforms (Ghermandi et al., 2022), resulting CES classes can diverge from established typologies such as CICES (Haines-Young, 2023). By grounding classification directly in human-understandable language prompts, CLIP-based methods provide more transparent and conceptually aligned CES outputs, strengthening the link between model inference and CES frameworks.

These conceptual advances are reflected in the empirical performance of our models. The CLIP-BRF-ZS pipeline achieved 88% overall accuracy, including strong performance on the "other'' class (F1-other = 0.91), using only 120 labeled examples. This represents a major improvement in label efficiency relative to existing vision-based models with transfer learning, which typically require one to two orders of magnitude more annotated images. For instance, recCNNize, an InceptionResNetV2 model fine-tuned on 11,912 Flickr images to recognize 12 recreational activities and two non-activity classes, reached 71% accuracy when evaluated in the region where it was trained, and 60% accuracy when applied to a novel region (Winder et al., 2022). Similarly, Cardoso et al. (2022) fine-tuned ResNet152 models on approximately 8,500 labeled Flickr images, reporting 87% accuracy for distinguishing natural from human-made elements and 77% accuracy for classifying six broad CES categories. Lingua et al. (2022, 2023) developed a multi-stage CNN pipeline based on ResNet-152, training one network to distinguish CES-relevant from irrelevant images and additional networks to classify broader CES groups and seven individual recreational activities, requiring 35,000 labeled images to achieve an accuracy of 85%. Across comparable recreational classes, our model achieved higher F1-scores (e.g., biking = 0.88 vs 0.78, camping = 0.88 vs. 0.70, hiking = 0.85 vs 0.82, wildlife viewing = 0.88 vs. 0.87) while using less than 0.4% of the labeled data used in prior studies.

Beyond improvements in label efficiency, the hybrid CLIP-BRF-ZS architecture effectively addresses the open-set challenges inherent in social media data. While zero-shot CLIP alone exhibited strong CES precision, it tended to over-reject valid CES images due to conservative cosine-similarity thresholding, and fully supervised CLIP-MRF performed poorly under low-label conditions. Integrating a lightweight binary filter allowed the subsequent zero-shot classifier to operate on cleaner inputs, improving recall and overall robustness. This targeted pre-filtering step yielded clear gains in open-set robustness: the CLIP-BRF-ZS model achieved a 0.91 F1-score for the "other'' class, compared with 0.68 in Lingua et al. (2022) and 0.46-0.61 in Winder et al. (2022). Notably, this improvement was achieved with only 60 CES-relevant and 60 CES-irrelevant training images, demonstrating that minimal supervision applied to CLIP's pretrained image embeddings can meaningfully enhance reliability under real-world conditions. Combined with its strong zero-shot CES performance, the hybrid approach preserves scalability and interpretability while offering substantially more robust open-set rejection.

4.2 The Role of Prompt Design

Building on CLIP's semantic flexibility, our results confirm that prompt design is a key determinant of zero-shot classification performance (Yong et al., 2023; Malla et al., 2025). Consistent with prior work showing that knowledge-embedded prompts outperform generic formulations (Yong et al., 2023; Shao et al., 2024), we found that context-rich prompts broadened the semantic scope of a category and improved recall—particularly for visually diffuse CES types such as wildlife viewing and landscape aesthetics. Because these categories vary widely in composition and may contain subtle cues, adding contextual elements effectively helped CLIP recognize a more diverse set of relevant images.

However, detailed prompts did not uniformly improve performance. For visually distinctive, action-oriented activities (e.g., fishing, camping), longer descriptions introduced semantic overlap with other outdoor activities and reduced inter-class separability. This aligns with evidence that CLIP's text encoder is tuned to short, caption-like descriptions emphasizing dominant objects or actions (Zhang et al., 2024) and that uniform prompt structures may be suboptimal across classes (Wen et al., 2025). To address this, we adopted a mixed-prompt strategy that pairs concise, action-centered prompts for distinctive activities with more descriptive prompts for abstract or context-dependent CES categories. This approach increased overall zero-shot accuracy by approximately five percentage points, demonstrating that effective prompt design requires balancing semantic detail with inter-class distinctiveness. Overall, these findings show that tailoring the semantic richness of prompts to each CES class, rather than uniformly increasing prompt detail, is effective for improving zero-shot CES classification.

4.3 Implications for Landscape Management

The mapped CES flow patterns reveal that cultural benefits in Florida arise from a heterogeneous landscape mosaic that extends well beyond the boundaries of formal parks. The balanced yet moderate CES flows in state and national parks suggests their role as multifunctional but not dominant service benefiting areas. Instead, urban greenspaces, protected areas, and a wide range of NWLs, including waters, wetlands, croplands and rangelands, supported a large share of CES flows. Aquatic and coastal environments emerged as important locations for CES engagement, consistent with studies showing that proximity to water and shoreline amplifies outdoor recreational and aesthetic value (Ghermandi et al., 2009; Sander and Zhao, 2015). These systems supported high levels of beach recreation, fishing, boating, and scenic enjoyment, underscoring the centrality of Florida's blue infrastructures to human–nature interactions. Working landscapes also played a meaningful role. Despite being underrepresented in CES research (Liu et al., 2019; Petway et al., 2020), cultivated croplands exhibited the highest per-user intensities, driven by activities including camping and fishing, while rangelands supported horseback riding and other forms of outdoor recreation. These patterns align with growing evidence that agricultural landscapes provide more than food production, contributing cultural and other ecosystem services that enhance human wellbeing (Swinton et al., 2007; van Berkel and Verburg, 2014). Policies that incentivize agroecological stewardship or expand recreational access can enhance these multifunctional benefits and sustain the cultural value of working lands alongside production goals (Philip Robertson et al., 2014; Plieninger et al., 2015). Together, our findings highlight the importance of landscape-scale planning that integrates parks, urban greenspaces, protected areas and other NWLs into coordinated conservation and recreation planning. The diversity of landscapes supporting CES flows illustrates that maintaining and enhancing a broad portfolio of NWLs can help sustain and expand opportunities for CES delivery across Florida, and our

interactive mapping platform provides a practical interface for visualizing these patterns and supporting landscape and spatial planning.

4.4 Limitation and Future Directions

Several limitations should be considered when interpreting our findings. First, while Flickr remains one of the most widely used social media platforms for CES research due to its open data accessibility and its ability to serve as a good proxy for recreational visitation in natural areas (Ghermandi, 2022), overall platform activity has declined in recent years, and its user base is not fully representative of the general population (Leppämäki et al., 2025). Although this study focused on pre-COVID periods and per-user engagement in our dataset remained relatively stable over time (Figure S.2), future work should validate Flickr-based CES flow estimates against survey-based visitation or anonymous mobility datasets to better assess their representativeness of on-site visits across NWLs.

Model-related limitations also warrant attention. Recent advances in LLMs have enabled more sophisticated semantic interpretation, inference, and reasoning (Chang et al., 2023). While large proprietary multimodal LLMs (e.g., GPT-5 Vision, Gemini 3, and GPT-4o) may offer enhanced semantic reasoning to CES (Khaldi et al., 2025), their closed-source nature, high inference cost, and limited accessibility constrain their suitability for reproducible large-scale CES mapping. To ensure transparency and reproducibility, we focused on open-source VLMs that can be locally deployed. In a related study, we further evaluated three open-source billion-parameter VLMs on 3.42 million Flickr images from the southeastern U.S. across the same CES categories (Liao et al., 2025). Qwen2.5-VL-7B achieved the highest zero-shot accuracy (90%), followed by Gemma-3-4B (79%) and Aya Vision-8B (73%). However, these larger models were computationally demanding, processing only 1–2 images per second and requiring large-memory GPUs (e.g., 180 GB on a NVIDIA B200). In contrast, the MobileCLIP-S1 model used in this study (~85 million parameters) achieved nearly comparable accuracy (88%) at a fraction of the computational cost, processing roughly 10–20 images per second on a single GPU. CLIP-based models are thus well-suited for regional- to continental-scale CES mapping, offering high throughput without substantial API or computational costs. Future work should evaluate the extent to which CLIP-based models can accurately identify CES categories that involve deeper semantic or symbolic meaning, such as cultural heritage or spiritual inspiration. In addition, advances in prompt engineering and lightweight adaptation techniques, including conditional prompt learning (Zhou et al., 2022b) and CLIP-Adapter for efficient few-shot transfer learning (Gao et al., 2024) may further enhance CLIP's capacity to distinguish CES categories with subtle or abstract visual expressions.

Finally, the present study focused solely on image analysis and did not incorporate textual metadata, which represents an emerging opportunity for multimodal CES assessment. Flickr's user-generated metadata, including titles, tags and descriptions, can provide useful semantic information for understanding people's perceptions and preferences. Recent work has used this textual information for sentiment analysis (Brindley et al., 2019; Fox et al., 2021). Havinga et al. 2024 integrated Flickr image–text pairs to examine links between people's positive experiences of nature and CES supply measures, demonstrating Flickr's potential of multimodal CES inference. Additional studies show that data from diverse modalities, including remote sensing imagery (Havinga et al., 2021b; Karasov et al., 2022), climate data (Manley and Egoh, 2022), and 3D point clouds (Hu et al., 2022), can further enrich CES interpretation. Moving forward, integrating imagery, textual information, and environmental data through multimodal approaches would

advance CES assessments from identifying where cultural benefits occur toward building a deeper understanding of how people perceive and experience them.

5. Conclusion

This study demonstrates that CLIP-based VLMs provide a scalable, label-efficient, and transferable framework for classifying CES from social media imagery and mapping CES flows across heterogeneous landscapes in Florida. Using only 120 labeled images for development, we show that MobileCLIP-S1 delivers strong zero-shot performance across 12 CES classes, and that class-specific prompt engineering is essential for optimizing CLIP's text–image alignment. Mixed prompts tailored to each CES class increased closed-set accuracy to 97%, with detailed descriptions improving recognition of visually or semantically overlapping categories, while concise prompts remained effective for more distinctive activities. These results highlight how small, strategically curated prompt sets can substantially enhance CES classification accuracy with minimal annotation effort.

A central methodological contribution of this study is the evaluation of three CLIP-based modeling pipelines under realistic open-set conditions, where a considerable fraction of Flickr images are irrelevant to CES. The hybrid CLIP-BRF-ZS approach with a lightweight binary classifier filtering irrelevant images prior to zero-shot inference achieved the highest overall accuracy (88%) and macro F1-score (0.88), along with robust rejection of irrelevant content (F1-other = 0.91). This pipeline also substantially improved recall for challenging CES classes such as fishing and landscape aesthetics. In contrast, CLIP-ZS-TF's conservative thresholding led to systematic mis-rejection of valid CES imagery, while the fully supervised CLIP-MRF performed poorly under limited labeled data, especially for landscape aesthetics. Together, these findings show that modest supervision applied to frozen CLIP embeddings can enable robust and label-efficient CES classification at regional scales while addressing the open-set recognition challenges inherent in social media data.

Statewide CES flow maps reveal that cultural benefits in Florida arise from a broad and diverse landscape mosaic extending far beyond national and state parks. Outdoor recreation, wildlife viewing, and landscape aesthetics exhibited similar spatial patterns, with shared hotspots along both coasts and major inland greenspaces. While protected areas played dominant roles for wildlife viewing, accessible urban greenspaces and other NWLs together supported roughly half of outdoor recreation and landscape aesthetics. Aquatic, wetland, and coastal environments emerged as major CES benefiting areas. Working lands, including croplands and pasture/hay, also contributed meaningfully, supporting outdoor recreation such as camping, fishing, and horseback riding, highlighting the multifunctionality of Florida's agricultural landscapes.

Overall, this study demonstrates that CLIP-based foundation VLMs advance CES assessment beyond reliance on extensive labeled datasets and conventional vision-based models, enabling a language-grounded, label-efficient paradigm for reproducible and spatially explicit analysis of crowdsourced social media imagery. By integrating label-efficiency, transparent text-guided reasoning, and strong open-set robustness, our framework provides a scalable pathway for fine-resolution CES flow quantification and mapping. The resulting statewide products and interactive mapping platform offer actionable tools for identifying CES hotspots and the diverse natural and working landscapes that sustain human–nature interactions.

Acknowledgment
This research was supported by USDA NIFA Hatch grant No. FLA-AGR-006393, the University of Florida UF/IFAS Early Career Seed Grant No. P00133052, and startup funds.

References

Bagstad, K.J., Johnson, G.W., Voigt, B., Villa, F., 2013. Spatial dynamics of ecosystem service flows: A comprehensive approach to quantifying actual services. Ecosyst. Serv., Special Issue on Mapping and Modelling Ecosystem Services 4, 117–125. https://doi.org/10.1016/j.ecoser.2012.07.012

Baró, F., Palomo, I., Zulian, G., Vizcaino, P., Haase, D., Gómez-Baggethun, E., 2016. Mapping ecosystem service capacity, flow and demand for landscape and urban planning: A case study in the Barcelona metropolitan region. Land Use Policy 57, 405–417. https://doi.org/10.1016/j.landusepol.2016.06.006

Bishop, C.M., Nasrabadi, N.M., 2006. Pattern recognition and machine learning. Springer.

Bordes, F., Pang, R.Y., Ajay, A., Li, A.C., Bardes, A., Petryk, S., Mañas, O., Lin, Z., Mahmoud, A., Jayaraman, B., Ibrahim, M., Hall, M., Xiong, Y., Lebensold, J., Ross, C., Jayakumar, S., Guo, C., Bouchacourt, D., Al-Tahan, H., Padthe, K., Sharma, V., Xu, H., Tan, X.E., Richards, M., Lavoie, S., Astolfi, P., Hemmat, R.A., Chen, J., Tirumala, K., Assouel, R., Moayeri, M., Talattof, A., Chaudhuri, K., Liu, Z., Chen, X., Garrido, Q., Ullrich, K., Agrawal, A., Saenko, K., Celikyilmaz, A., Chandra, V., 2024. An Introduction to Vision-Language Modeling. https://doi.org/10.48550/arXiv.2405.17247

Breiman, L., 2001. Random Forests. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324

Brindley, P., Cameron, R.W., Ersoy, E., Jorgensen, A., Maheswaran, R., 2019. Is more always better? Exploring field survey and social media indicators of quality of urban greenspace, in relation to health. Urban For. Urban Green. 39, 45–54. https://doi.org/10.1016/j.ufug.2019.01.015

Burkhard, B., Kandziora, M., Hou, Y., Müller, F., 2014. Ecosystem service potentials, flows and demands-concepts for spatial localisation, indication and quantification. Landsc. Online 34–34. https://doi.org/10.3097/LO.201434

Burkhard, B., Maes, J., 2017. Mapping Ecosystem Services. Adv. Books 1, e12837. https://doi.org/10.3897/ab.e12837

Cao, H., Wang, M., Su, S., Kang, M., 2022. Explicit quantification of coastal cultural ecosystem services: A novel approach based on the content and sentimental analysis of social media. Ecol. Indic. 137, 108756. https://doi.org/10.1016/j.ecolind.2022.108756

Cardoso, A.S., Renna, F., Moreno-Llorca, R., Alcaraz-Segura, D., Tabik, S., Ladle, R.J., Vaz, A.S., 2022. Classifying the content of social media images to support cultural ecosystem service assessments using deep learning models. Ecosyst. Serv. 54, 101410.

Chai-allah, A., Hermes, J., La Foye, A.D., Venter, Z.S., Joly, F., Brunschwig, G., Bimonte, S., Fox, N., 2025. Assessing recreationists' preferences of the landscape and species using crowdsourced images and machine learning. Landsc. Urban Plan. 257, 105315. https://doi.org/10.1016/j.landurbplan.2025.105315

Chang, Yupeng, Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., Ye, W., Zhang, Y., Chang, Yi, Yu, P.S., Yang, Q., Xie, X., 2023. A Survey on Evaluation of Large Language Models. https://doi.org/10.48550/arXiv.2307.03109

Chen, S., Wang, X., Liu, T., Xie, M., Lin, Q., 2025. Using geo-data and social media images to explore the supply and demand of cultural ecosystem services for terraces in China. Ecosyst. Serv. 76, 101778. https://doi.org/10.1016/j.ecoser.2025.101778

Cheng, X., Van Damme, S., Li, L., Uyttenhove, P., 2019. Evaluation of cultural ecosystem services: A review of methods. Ecosyst. Serv. 37, 100925. https://doi.org/10.1016/j.ecoser.2019.100925

Cherti, M., Beaumont, R., Wightman, R., Wortsman, M., Ilharco, G., Gordon, C., Schuhmann, C., Schmidt, L., Jitsev, J., 2023. Reproducible scaling laws for contrastive language-image learning, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 2818–2829.

Comalada, F., Acuña, V., Garcia, X., 2025. Modelling cultural ecosystem services of river landscapes in the Iberian Peninsula with deep learning and social media images. J. Environ. Manage. 394, 127667.

Costadone, L., Balzan, M.V., 2023. Characterizing nature-based recreation preferences in a Mediterranean small island environment using crowdsourced data. Ecosyst. People 19, 2274594. https://doi.org/10.1080/26395916.2023.2274594

Daniel, T.C., Muhar, A., Arnberger, A., Aznar, O., Boyd, J.W., Chan, K.M.A., Costanza, R., Elmqvist, T., Flint, C.G., Gobster, P.H., Grêt-Regamey, A., Lave, R., Muhar, S., Penker, M., Ribe, R.G., Schauppenlehner, T., Sikor, T., Soloviy, I., Spierenburg, M., Taczanowska, K., Tam, J., von der Dunk, A., 2012. Contributions of cultural services to the ecosystem services agenda. Proc. Natl. Acad. Sci. 109, 8812–8819. https://doi.org/10.1073/pnas.1114773109

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. ImageNet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition. Presented at the 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255. https://doi.org/10.1109/CVPR.2009.5206848

Dosovitskiy, A., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. ArXiv Prepr. ArXiv201011929.

Edens, B., Maes, J., Hein, L., Obst, C., Siikamaki, J., Schenau, S., Javorsek, M., Chow, J., Chan, J.Y., Steurer, A., Alfieri, A., 2022. Establishing the SEEA Ecosystem Accounting as a global standard. Ecosyst. Serv. 54, 101413. https://doi.org/10.1016/j.ecoser.2022.101413

Egarter Vigl, L., Marsoner, T., Giombini, V., Pecher, C., Simion, H., Stemle, E., Tasser, E., Depellegrin, D., 2021. Harnessing artificial intelligence technology and social media data to support Cultural Ecosystem Service assessments. People Nat. 3, 673–685. https://doi.org/10.1002/pan3.10199

FDEP, 2025. Florida State Parks Boundaries. https://geodata.dep.state.fl.us/datasets/florida-state-parks-boundaries/explore

Fisher, B., Turner, R.K., Morling, P., 2009. Defining and classifying ecosystem services for decision making. Ecol. Econ. 68, 643–653. https://doi.org/10.1016/j.ecolecon.2008.09.014

Fox, N., Graham, L.J., Eigenbrod, F., Bullock, J.M., Parks, K.E., 2021. Enriching social media data allows a more robust representation of cultural ecosystem services. Ecosyst. Serv. 50, 101328. https://doi.org/10.1016/j.ecoser.2021.101328

Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y., 2024. CLIP-Adapter: Better Vision-Language Models with Feature Adapters. Int. J. Comput. Vis. 132, 581–595. https://doi.org/10.1007/s11263-023-01891-x

Ghermandi, A., 2022. Geolocated social media data counts as a proxy for recreational visits in natural areas: A meta-analysis. J. Environ. Manage. 317, 115325. https://doi.org/10.1016/j.jenvman.2022.115325
Ghermandi, A., Depietri, Y., Sinclair, M., 2022. In the AI of the beholder: A comparative analysis of computer vision-assisted characterizations of human-nature interactions in urban green spaces. Landsc. Urban Plan. 217, 104261. https://doi.org/10.1016/j.landurbplan.2021.104261
Ghermandi, A., Langemeyer, J., Van Berkel, D., Calcagni, F., Depietri, Y., Egarter Vigl, L., Fox, N., Havinga, I., Jäger, H., Kaiser, N., Karasov, O., McPhearson, T., Podschun, S., Ruiz-Frau, A., Sinclair, M., Venohr, M., Wood, S.A., 2023. Social media data for environmental sustainability: A critical review of opportunities, threats, and ethical use. One Earth 6, 236–250. https://doi.org/10.1016/j.oneear.2023.02.008
Ghermandi, A., Nunes, P.A.L.D., Portela, R., Rao, N., Teelucksingh, S.S., 2009. Recreational, Cultural and Aesthetic Services from Estuarine and Coastal Ecosystems (Fondazione Eni Enrico Mattei Working Papers), SD. https://doi.org/10.22004/ag.econ.56214
Ghermandi, A., Sinclair, M., 2019. Passive crowdsourcing of social media in environmental research: A systematic map. Glob. Environ. Change 55, 36–47. https://doi.org/10.1016/j.gloenvcha.2019.02.003
Goldspiel, H., Barr, B., Badding, J., Kuehn, D., 2023. Snapshots of Nature-Based Recreation Across Rural Landscapes: Insights from Geotagged Photographs in the Northeastern United States. Environ. Manage. 71, 234–248. https://doi.org/10.1007/s00267-022-01728-2
Haines-Young, R., 2023. Common International Classification of Ecosystem Services (CICES) V5.2 and Guidance on the Application of the Revised Structure.
Havinga, I., Bogaart, P.W., Hein, L., Tuia, D., 2020. Defining and spatially modelling cultural ecosystem services using crowdsourced data. Ecosyst. Serv. 43, 101091. https://doi.org/10.1016/j.ecoser.2020.101091
Havinga, I., Marcos, D., Bogaart, P., Massimino, D., Hein, L., Tuia, D., 2023. Social media and deep learning reveal specific cultural preferences for biodiversity. People Nat. 5, 981–998. https://doi.org/10.1002/pan3.10466
Havinga, I., Marcos, D., Bogaart, P., Tuia, D., Hein, L., 2024. Understanding the sentiment associated with cultural ecosystem services using images and text from social media. Ecosyst. Serv. 65, 101581. https://doi.org/10.1016/j.ecoser.2023.101581
Havinga, I., Marcos, D., Bogaart, P.W., Hein, L., Tuia, D., 2021a. Social media and deep learning capture the aesthetic quality of the landscape. Sci. Rep. 11, 20000. https://doi.org/10.1038/s41598-021-99282-0
Havinga, I., Marcos, D., Bogaart, P.W., Hein, L., Tuia, D., 2021b. Geo-Data for Mapping Scenic Beauty: Exploring the Potential of Remote Sensing and Social Media, in: 2021 IEEE International Geoscience and Remote Sensing Symposium IGARSS. Presented at the IGARSS 2021 - 2021 IEEE International Geoscience and Remote Sensing Symposium, IEEE, Brussels, Belgium, pp. 188–191. https://doi.org/10.1109/IGARSS47720.2021.9553417
He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 770–778.

Hu, T., Wei, D., Su, Y., Wang, X., Zhang, J., Sun, X., Liu, Y., Guo, Q., 2022. Quantifying the shape of urban street trees and evaluating its influence on their aesthetic functions based on mobile lidar data. ISPRS J. Photogramm. Remote Sens. 184, 203–214.

Huai, S., Chen, F., Liu, S., Canters, F., Van de Voorde, T., 2022. Using social media photos and computer vision to assess cultural ecosystem services and landscape features in urban parks. Ecosyst. Serv. 57, 101475.

Huynh, L.T.M., Gasparatos, A., Su, J., Dam Lam, R., Grant, E.I., Fukushi, K., 2022. Linking the nonmaterial dimensions of human-nature relations and human well-being through cultural ecosystem services. Sci. Adv. 8, eabn8042. https://doi.org/10.1126/sciadv.abn8042

Ilharco, G., Wortsman, M., Wightman, R., Gordon, C., Carlini, N., Taori, R., Dave, A., Shankar, V., Namkoong, H., Miller, J., Hajishirzi, H., Farhadi, A., Schmidt, L., 2025. OpenCLIP. https://doi.org/10.5281/zenodo.14913962

Kang, R., Song, Y., Gkioxari, G., Perona, P., 2025. Is CLIP ideal? No. Can we fix it? Yes! https://doi.org/10.48550/arXiv.2503.08723

Karasov, O., Heremans, S., Külvik, M., Domnich, A., Burdun, I., Kull, A., Helm, A., Uuemaa, E., 2022. Beyond land cover: How integrated remote sensing and social media data analysis facilitates assessment of cultural ecosystem services. Ecosyst. Serv. 53, 101391.

Keeler, B.L., Wood, S.A., Polasky, S., Kling, C., Filstrup, C.T., Downing, J.A., 2015. Recreational demand for clean water: evidence from geotagged photographs by visitors to lakes. Front. Ecol. Environ. 13, 76–81. https://doi.org/10.1890/140124

Khaldi, R., Alcaraz-Segura, D., Sánchez-Herrera, I., Martinez-Lopez, J., Navarro, C.J., Tabik, S., 2025. Exploring Social Media Image Categorization Using Large Models with Different Adaptation Methods: A Case Study on Cultural Nature's Contributions to People. https://doi.org/10.48550/arXiv.2410.00275

Kosanic, A., Petzold, J., 2020. A systematic review of cultural ecosystem services and human wellbeing. Ecosyst. Serv. 45, 101168. https://doi.org/10.1016/j.ecoser.2020.101168

Koylu, C., Zhao, C., Shao, W., 2019. Deep Neural Networks and Kernel Density Estimation for Detecting Human Activity Patterns from Geo-Tagged Images: A Case Study of Birdwatching on Flickr. ISPRS Int. J. Geo-Inf. 8, 45. https://doi.org/10.3390/ijgi8010045

Langemeyer, J., Calcagni, F., Baró, F., 2018. Mapping the intangible: Using geolocated social media data to examine landscape aesthetics. Land Use Policy 77, 542–552. https://doi.org/10.1016/j.landusepol.2018.05.049

Langemeyer, J., Ghermandi, A., Keeler, B., van Berkel, D., 2023. The future of crowd-sourced cultural ecosystem services assessments. Ecosyst. Serv. 60, 101518. https://doi.org/10.1016/j.ecoser.2023.101518

Lee, H., Seo, B., Koellner, T., Lautenbach, S., 2019. Mapping cultural ecosystem services 2.0 – Potential and shortcomings from unlabeled crowd sourced images. Ecol. Indic. 96, 505–515. https://doi.org/10.1016/j.ecolind.2018.08.035

Leppämäki, T., Heikinheimo, V., Eklund, J., Hausmann, A., Toivonen, T., 2025. The rise and fall of the social media platform Flickr: Implications for nature recreation research. J. Outdoor Recreat. Tour. 50, 100880. https://doi.org/10.1016/j.jort.2025.100880

Levering, A., Marcos, D., Jacobs, N., Tuia, D., 2024. Prompt-guided and multimodal landscape scenicness assessments with vision-language models. PLOS ONE 19, e0307083. https://doi.org/10.1371/journal.pone.0307083

Li, J., Wang, Y., 2023. Ecosystem services assessment from capacity to flow: A review. Trans. Earth Environ. Sustain. 1, 80–93. https://doi.org/10.1177/2754124X221141991
Liao, H.-Y., Zhao, C., Song, J., Shao, W., 2025. Mapping Cultural Ecosystem Services Using One-Shot In-Context Learning with Multimodal Large Language Models, in: Proceedings of the 33rd ACM International Conference on Advances in Geographic Information Systems (SIGSPATIAL '25). Association for Computing Machinery, Minneapolis, MN, USA, p. 4. https://doi.org/10.1145/3748636.3764178
Lingua, F., Coops, N.C., Griess, V.C., 2023. Assessing forest recreational potential from social media data and remote sensing technologies data. Ecol. Indic. 149, 110165. https://doi.org/10.1016/j.ecolind.2023.110165
Lingua, F., Coops, N.C., Griess, V.C., 2022. Valuing cultural ecosystem services combining deep learning and benefit transfer approach. Ecosyst. Serv. 58, 101487.
Liu, L., Yang, L., Yang, F., Chen, F., Xu, F., 2024. CLIP-Driven Few-Shot Species-Recognition Method for Integrating Geographic Information. Remote Sens. 16, 2238. https://doi.org/10.3390/rs16122238
Liu, W., Wang, J., Li, C., Chen, B., Sun, Y., 2019. Using Bibliometric Analysis to Understand the Recent Progress in Agroecosystem Services Research. Ecol. Econ. 156, 293–305. https://doi.org/10.1016/j.ecolecon.2018.09.001
Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., Xie, S., 2022. A convnet for the 2020s, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11976–11986.
Luo, H., Zhang, Z., Zhu, Q., Ameur, N.E.H.B., Liu, X., Ding, F., Cai, Y., 2025. Using large language models to investigate cultural ecosystem services perceptions: A few-shot and prompt method. Landsc. Urban Plan. 258, 105323.
MA, 2005. Ecosystems and Human Well-being: Synthesis. Island Press, Washington, DC.
Maes, J., Egoh, B., Willemen, L., Liquete, C., Vihervaara, P., Schägner, J.P., Grizzetti, B., Drakou, E.G., Notte, A.L., Zulian, G., Bouraoui, F., Luisa Paracchini, M., Braat, L., Bidoglio, G., 2012. Mapping ecosystem services for policy support and decision making in the European Union. Ecosyst. Serv. 1, 31–39. https://doi.org/10.1016/j.ecoser.2012.06.004
Malla, H.J., Bazli, M., Arashpour, M., 2025. Enhancing waste recognition with vision-language models: A prompt engineering approach for a scalable solution. Waste Manag. 204, 114939.
Manley, K., Egoh, B.N., 2022. Mapping and modeling the impact of climate change on recreational ecosystem services using machine learning and big data. Environ. Res. Lett. 17, 054025. https://doi.org/10.1088/1748-9326/ac65a3
Manley, K., Nyelele, C., Egoh, B.N., 2022. A review of machine learning and big data applications in addressing ecosystem service research gaps. Ecosyst. Serv. 57, 101478. https://doi.org/10.1016/j.ecoser.2022.101478
Martínez Pastur, G., Peri, P.L., Lencinas, M.V., García-Llorente, M., Martín-López, B., 2016. Spatial patterns of cultural ecosystem services provision in Southern Patagonia. Landsc. Ecol. 31, 383–399. https://doi.org/10.1007/s10980-015-0254-9
Miller, D., Sünderhauf, N., Kenna, A., Mason, K., 2025. Open-Set Recognition in the Age of Vision-Language Models, in: Leonardis, A., Ricci, E., Roth, S., Russakovsky, O., Sattler, T., Varol, G. (Eds.), Computer Vision – ECCV 2024. Springer Nature Switzerland, Cham, pp. 1–18. https://doi.org/10.1007/978-3-031-72946-1_1

Mitten, D., Overholt, J.R., Haynes, F.I., D'Amore, C.C., Ady, J.C., 2018. Hiking: A Low-Cost, Accessible Intervention to Promote Health Benefits. Am. J. Lifestyle Med. 12, 302–310. https://doi.org/10.1177/1559827616658229

Naidoo, R., Balmford, A., Costanza, R., Fisher, B., Green, R.E., Lehner, B., Malcolm, T.R., Ricketts, T.H., 2008. Global mapping of ecosystem services and conservation priorities. Proc. Natl. Acad. Sci. 105, 9495–9500. https://doi.org/10.1073/pnas.0707823105

NPS Land Resources Division, 2025. Administrative Boundaries of National Park System Units - National Geospatial Data Asset (NGDA) NPS National Parks Dataset. NPS - Land Resources Division. United States and Territories. https://irma.nps.gov/DataStore/Reference/Profile/2314463

Oteros-Rozas, E., Martín-López, B., Fagerholm, N., Bieling, C., Plieninger, T., 2018. Using social media photos to explore the relation between cultural ecosystem services and landscape features across five European sites. Ecol. Indic. 94, 74–86. https://doi.org/10.1016/j.ecolind.2017.02.009

Paracchini, M.L., Zulian, G., Kopperoinen, L., Maes, J., Schägner, J.P., Termansen, M., Zandersen, M., Perez-Soba, M., Scholefield, P.A., Bidoglio, G., 2014. Mapping cultural ecosystem services: A framework to assess the potential for outdoor recreation across the EU. Ecol. Indic. 45, 371–385. https://doi.org/10.1016/j.ecolind.2014.04.018

Park, J., Lee, J., Song, J., Yu, S., Jung, D., Yoon, S., 2025. Know "No'' Better: A Data-Driven Approach for Enhancing Negation Awareness in CLIP. https://doi.org/10.48550/arXiv.2501.10913

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., 2019. Pytorch: An imperative style, high-performance deep learning library. Adv. Neural Inf. Process. Syst. 32.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: Machine Learning in Python. J. Mach. Learn. Res. 12, 2825–2830.

Petway, J.R., Lin, Y.-P., Wunderlich, R.F., 2020. A place-based approach to agricultural nonmaterial intangible cultural ecosystem service values. Sustainability 12, 699.

Philip Robertson, G., Gross, K.L., Hamilton, S.K., Landis, D.A., Schmidt, T.M., Snapp, S.S., Swinton, S.M., 2014. Farming for Ecosystem Services: An Ecological Approach to Production Agriculture. BioScience 64, 404–415. https://doi.org/10.1093/biosci/biu037

Plieninger, T., Bieling, C., Fagerholm, N., Byg, A., Hartel, T., Hurley, P., López-Santiago, C.A., Nagabhatla, N., Oteros-Rozas, E., Raymond, C.M., van der Horst, D., Huntsinger, L., 2015. The role of cultural ecosystem services in landscape management and planning. Curr. Opin. Environ. Sustain., Open Issue 14, 28–33. https://doi.org/10.1016/j.cosust.2015.02.006

Plieninger, T., Dijks, S., Oteros-Rozas, E., Bieling, C., 2013. Assessing, mapping, and quantifying cultural ecosystem services at community level. Land Use Policy 33, 118–129. https://doi.org/10.1016/j.landusepol.2012.12.013

Quantmeyer, V., Mosteiro, P., Gatt, A., 2024. How and where does CLIP process negation? https://doi.org/10.48550/arXiv.2407.10488

Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., 2021. Learning transferable visual models from natural language supervision, in: International Conference on Machine Learning. PmLR, pp. 8748–8763.

Richards, D.R., Friess, D.A., 2015. A rapid indicator of cultural ecosystem service usage at a fine spatial scale: Content analysis of social media photographs. Ecol. Indic. 53, 187–195. https://doi.org/10.1016/j.ecolind.2015.01.034
Richards, D.R., Lavorel, S., 2022. Integrating social media data and machine learning to analyse scenarios of landscape appreciation. Ecosyst. Serv. 55, 101422. https://doi.org/10.1016/j.ecoser.2022.101422
Richards, D.R., Tunçer, B., 2018. Using image recognition to automate assessment of cultural ecosystem services from social media photographs. Ecosyst. Serv., Assessment and Valuation of Recreational Ecosystem Services 31, 318–325. https://doi.org/10.1016/j.ecoser.2017.09.004
Rosenberger, R.S., White, E.M., Kline, J.D., Cvitanovich, C., 2017. Recreation economic values for estimating outdoor recreation economic benefits from the National Forest System.
Russe, M.F., Reisert, M., Bamberg, F., Rau, A., 2024. Improving the use of LLMs in radiology through prompt engineering: from precision prompts to zero-shot learning. RöFo - Fortschritte Auf Dem Geb. Röntgenstrahlen Bildgeb. Verfahr. 196, 1166–1170. https://doi.org/10.1055/a-2264-5631
Sander, H.A., Zhao, C., 2015. Urban green and blue: Who values what and where? Land Use Policy 42, 194–209. https://doi.org/10.1016/j.landusepol.2014.07.021
Schirpke, U., Ghermandi, A., Sinclair, M., Van Berkel, D., Fox, N., Vargas, L., Willemen, L., 2023. Emerging technologies for assessing ecosystem services: A synthesis of opportunities and challenges. Ecosyst. Serv. 63, 101558. https://doi.org/10.1016/j.ecoser.2023.101558
Schröter, M., Barton, D.N., Remme, R.P., Hein, L., 2014. Accounting for capacity and flow of ecosystem services: A conceptual model and a case study for Telemark, Norway. Ecol. Indic. 36, 539–551. https://doi.org/10.1016/j.ecolind.2013.09.018
Scowen, M., Athanasiadis, I.N., Bullock, J.M., Eigenbrod, F., Willcock, S., 2021. The current and future uses of machine learning in ecosystem service research. Sci. Total Environ. 799, 149263. https://doi.org/10.1016/j.scitotenv.2021.149263
Serna-Chavez, H.M., Schulp, C.J.E., van Bodegom, P.M., Bouten, W., Verburg, P.H., Davidson, M.D., 2014. A quantitative framework for assessing spatial flows of ecosystem services. Ecol. Indic. 39, 24–33. https://doi.org/10.1016/j.ecolind.2013.11.024
Sessions, C., Wood, S.A., Rabotyagov, S., Fisher, D.M., 2016. Measuring recreational visitation at U.S. National Parks with crowd-sourced photographs. J. Environ. Manage. 183, 703–711. https://doi.org/10.1016/j.jenvman.2016.09.018
Shao, Y., Li, P., Yu, L., Wang, S., Ai, R., 2024. CLIP-optimized prompt for zero-shot wild animal identification, in: Ninth International Workshop on Pattern Recognition. SPIE, p. 1339902.
Shtedritski, A., Rupprecht, C., Vedaldi, A., 2023. What does CLIP know about a red circle? Visual prompt engineering for VLMs. https://doi.org/10.48550/arXiv.2304.06712
Silverman, B.W., 1986. Density Estimation for Statistics and Data Analysis. Routledge, New York. https://doi.org/10.1201/9781315140919
Simonyan, K., Zisserman, A., 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. https://doi.org/10.48550/arXiv.1409.1556
Sinclair, M., Mayer, M., Woltering, M., Ghermandi, A., 2020. Using social media to estimate visitor provenance and patterns of recreation in Germany's national parks. J. Environ. Manage. 263, 110418. https://doi.org/10.1016/j.jenvman.2020.110418

Sonter, L.J., Watson, K.B., Wood, S.A., Ricketts, T.H., 2016. Spatial and Temporal Dynamics and Value of Nature-Based Recreation, Estimated via Social Media. PLOS ONE 11, e0162372. https://doi.org/10.1371/journal.pone.0162372

Sun, Q., Fang, Y., Wu, L., Wang, X., Cao, Y., 2023. EVA-CLIP: Improved Training Techniques for CLIP at Scale. https://doi.org/10.48550/arXiv.2303.15389

Swinton, S.M., Lupi, F., Robertson, G.P., Hamilton, S.K., 2007. Ecosystem services and agriculture: Cultivating agricultural ecosystems for diverse benefits. Ecol. Econ. 64, 245–252. https://doi.org/10.1016/j.ecolecon.2007.09.020

Tan, H.-Z., Zhou, Z., Li, Y., Guo, L.-Z., 2025. Vision-Language Model Selection and Reuse for Downstream Adaptation, in: Forty-Second International Conference on Machine Learning.

Torrey, L., Shavlik, J., 2010. Transfer Learning, in: Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques. IGI Global Scientific Publishing, pp. 242–264. https://doi.org/10.4018/978-1-60566-766-9.ch011

Trust for Public Land, 2025. ParkServe. https://www.tpl.org/parkserve

U.S. Census Bureau, 2020. Census Urban Areas. https://doi.org/https:%2520//www2.census.gov/geo/tiger/TIGER2023/UAC/

U.S. Climate Alliance, 2022. , Natural and Working Lands & Climate Action: A State Guide to Enhance the Sector's Contribution to State and National Climate Goals. Washington, D.C.

USGS, 2024. Annual NLCD Collection 1 Science Products (ver. 1.1, June 2025). https://doi.org/10.5066/P94UXNTS

USGS GAP, 2024. Protected Areas Database of the United States (PAD-US) 4. https://doi.org/10.5066/P96WBCHS.

Vallecillo, S., La Notte, A., Zulian, G., Ferrini, S., Maes, J., 2019. Ecosystem services accounts: Valuing the actual flow of nature-based recreation from ecosystems to people. Ecol. Model. 392, 196–211. https://doi.org/10.1016/j.ecolmodel.2018.09.023

van Berkel, D.B., Verburg, P.H., 2014. Spatial quantification and valuation of cultural ecosystem services in an agricultural landscape. Ecol. Indic. 37, 163–174. https://doi.org/10.1016/j.ecolind.2012.06.025

Vasu, P.K.A., Pouransari, H., Faghri, F., Vemulapalli, R., Tuzel, O., 2024. Mobileclip: Fast image-text models through multi-modal reinforced training, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15963–15974.

Villamagna, A.M., Angermeier, P.L., Bennett, E.M., 2013. Capacity, pressure, demand, and flow: A conceptual framework for analyzing ecosystem service provision and delivery. Ecol. Complex. 15, 114–121. https://doi.org/10.1016/j.ecocom.2013.07.004

Wang, L., Zheng, H., Chen, Y., Ouyang, Z., Hu, X., 2022. Systematic review of ecosystem services flow measurement: Main concepts, methods, applications and future directions. Ecosyst. Serv. 58, 101479. https://doi.org/10.1016/j.ecoser.2022.101479

Wen, C., Peng, Z., Huang, Y., Yang, X., Shen, W., 2025. Domain Generalization in CLIP via Learning with Diverse Text Prompts, in: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 9559–9569.

Willemen, L., Cottam, A.J., Drakou, E.G., Burgess, N.D., 2015. Using Social Media to Measure the Contribution of Red List Species to the Nature-Based Tourism Potential of African Protected Areas. PLOS ONE 10, e0129785. https://doi.org/10.1371/journal.pone.0129785

Winder, S.G., Lee, H., Seo, B., Lia, E.H., Wood, S.A., 2022. An open-source image classifier for characterizing recreational activities across landscapes. People Nat. 4, 1249–1262. https://doi.org/10.1002/pan3.10382

Wood, S.A., Guerry, A.D., Silver, J.M., Lacayo, M., 2013. Using social media to quantify nature-based tourism and recreation. Sci. Rep. 3, 2976. https://doi.org/10.1038/srep02976

Xu, H., Duan, J., Ren, M., Zhao, G., Liu, Z., 2025. Revealing youth-perceived cultural ecosystem services for high-density urban green space management: a deep learning spatial analysis of social media photographs from central Beijing. Landsc. Ecol. 40, 119. https://doi.org/10.1007/s10980-025-02115-y

Yee, T.B.L., Carrasco, L.R., 2024. Applying deep learning on social media to investigate cultural ecosystem services in protected areas worldwide. Sci. Rep. 14, 13700.

Yong, G., Jeon, K., Gil, D., Lee, G., 2023. Prompt engineering for zero-shot and few-shot defect detection and classification using a visual-language pretrained model. Comput.-Aided Civ. Infrastruct. Eng. 38, 1536–1554. https://doi.org/10.1111/mice.12954

You, W., Xu, H., Ren, M., Duan, J., Zhang, R., Kizos, T., 2025. Mapping multiple stakeholder-perceived cultural ecosystem services in coastal landscapes along the Maritime Silk Road. Landsc. Res. 1–21. https://doi.org/10.1080/01426397.2025.2501233

Zhang, B., Zhang, P., Dong, X., Zang, Y., Wang, J., 2024. Long-CLIP: Unlocking the Long-Text Capability of CLIP [WWW Document]. arXiv.org. URL https://arxiv.org/abs/2403.15378v3 (accessed 10.19.25).

Zhang, J., Huang, J., Jin, S., Lu, S., 2024. Vision-Language Models for Vision Tasks: A Survey. IEEE Trans. Pattern Anal. Mach. Intell. 46, 5625–5644. https://doi.org/10.1109/TPAMI.2024.3369699

Zhao, X., Lu, Y., Huang, W., Lin, G., 2024. Assessing and interpreting perceived park accessibility, usability and attractiveness through texts and images from social media. Sustain. Cities Soc. 112, 105619. https://doi.org/10.1016/j.scs.2024.105619

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., Torralba, A., 2018. Places: A 10 Million Image Database for Scene Recognition. IEEE Trans. Pattern Anal. Mach. Intell. 40, 1452–1464. https://doi.org/10.1109/TPAMI.2017.2723009

Zhou, K., Yang, J., Loy, C.C., Liu, Z., 2022a. Learning to Prompt for Vision-Language Models. Int. J. Comput. Vis. 130, 2337–2348. https://doi.org/10.1007/s11263-022-01653-1

Zhou, K., Yang, J., Loy, C.C., Liu, Z., 2022b. Conditional Prompt Learning for Vision-Language Models. Presented at the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16816–16825.

Supplementary

Table S.1. Class-wise cosine similarity thresholds derived from the prompt tuning dataset.

Activity	TF-Mean	TF-Minimum	TF-Maximum
Biking	0.215	0.180	0.261
Boating	0.205	0.136	0.282
Camping	0.251	0.206	0.283
Fishing	0.227	0.171	0.263
Hiking	0.252	0.198	0.287
Horseback Riding	0.272	0.176	0.306
Hunting	0.236	0.160	0.304
Shelling	0.251	0.187	0.309
Surfing	0.252	0.171	0.291
Swimming	0.231	0.188	0.262
Wildlife Viewing	0.187	0.155	0.228
Landscape Aesthetics	0.207	0.174	0.242

TF-Mean, TF-Minimum, and TF-Maximum correspond to the average, minimum, and maximum similarity scores, respectively, across all image–prompt pairs within each CES class.

Table S.2. Hyperparameters for Random Forest models.

Parameters	Grid Search	Binary RF	Multiclass RF
n_estimators	10, 50, 100, 200, 300	200	300
max_depth	None, 10, 20, 30	None	None
min_samples_split	2, 5, 10	2	10
min_samples_leaf	1, 2, 4	1	4
max_features	sqrt, log2, None	log2	log2
class_weight	balanced, balanced_subsample, None	balanced	balanced

Table S.3. Per-user CES flow intensity by CES classes and NWLs. Values represent mean $\overline{PUD}_{user,yr}$.

Class	Urban greenspaces	Non-urban greenspaces	State parks	National parks	Protected areas	Other NWLs
Biking	0.19	0.18	0.17	0.17	0.19	0.20
Boating	0.21	0.19	0.18	0.18	0.22	0.23
Camping	0.19	0.17	0.21	0.18	0.21	0.25
Fishing	0.20	0.17	0.17	0.19	0.22	0.24
Hiking	0.18	0.17	0.18	0.17	0.19	0.19
Horseback Riding	0.20	0.17	0.17	0.20	0.19	0.21
Hunting	0.19	0.00	0.18	0.25	0.20	0.25
Shelling	0.23	0.27	0.18	0.19	0.21	0.21
Surfing	0.20	0.18	0.18	0.19	0.20	0.20
Swimming	0.20	0.18	0.18	0.18	0.22	0.22
Wildlife Viewing	0.29	0.22	0.23	0.25	0.28	0.29
Landscape Aesthetics	0.24	0.19	0.20	0.21	0.21	0.21
Average	0.28	0.22	0.22	0.24	0.26	0.25

Table S.4. Per-user CES flow intensity by CES classes and LULC types. Values represent mean $\overline{PUD}_{user,yr}$.

Class	Wetlands	Water	Pasture /Hay	Barren	Forest	Cultivated Crops	Herbaceous	Shrubland
Biking	0.17	0.19	0.18	0.17	0.18	0.50	0.17	0.17
Boating	0.19	0.20	0.25	0.21	0.17	0.83	0.15	0.17
Camping	0.20	0.17	0.26	0.19	0.19	2.87	0.18	0.17
Fishing	0.20	0.19	0.20	0.19	0.17	2.90	0.17	0.15
Hiking	0.18	0.16	0.18	0.17	0.18	0.47	0.18	0.15
Horseback Riding	0.18	0.15	0.20	0.17	0.17	0.95	0.17	0.17
Hunting	0.19	0.23	0.19	0.18	0.18	1.53	0.13	0.17
Shelling	0.18	0.18	0.17	0.20	0.17	0.40	0.15	0.15
Surfing	0.18	0.20	0.24	0.20	0.18	0.50	0.25	0.17
Swimming	0.19	0.19	0.20	0.22	0.19	0.97	0.17	0.17
Wildlife Viewing	0.24	0.24	0.22	0.21	0.21	0.33	0.23	0.18
Landscape Aesthetics	0.19	0.20	0.18	0.21	0.20	0.26	0.17	0.17
Average	0.23	0.23	0.22	0.23	0.21	0.32	0.21	0.18

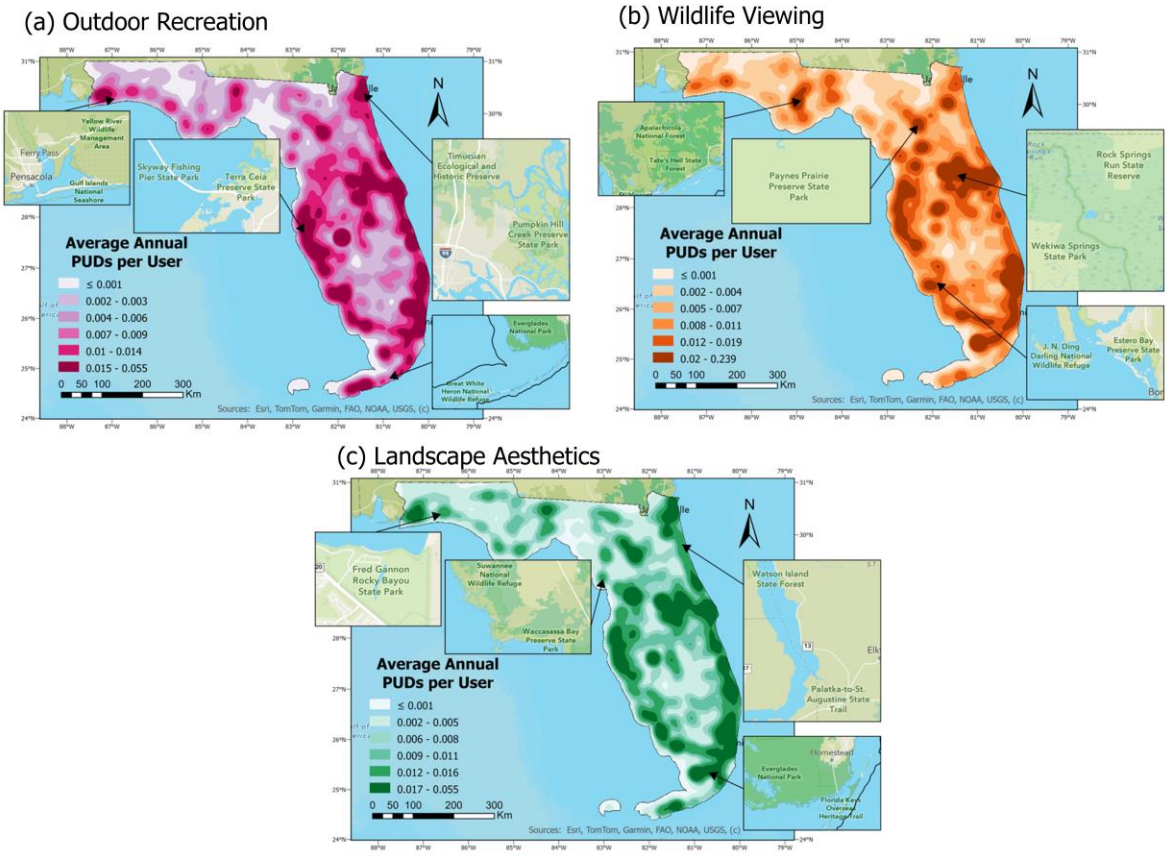

Figure S.1. Kernel density of average annual PUDs per user ($\overline{PUD}_{user,yr}$) at 1 km resolution across Florida for (a) outdoor recreation, (b) wildlife viewing, and (c) landscape aesthetics.

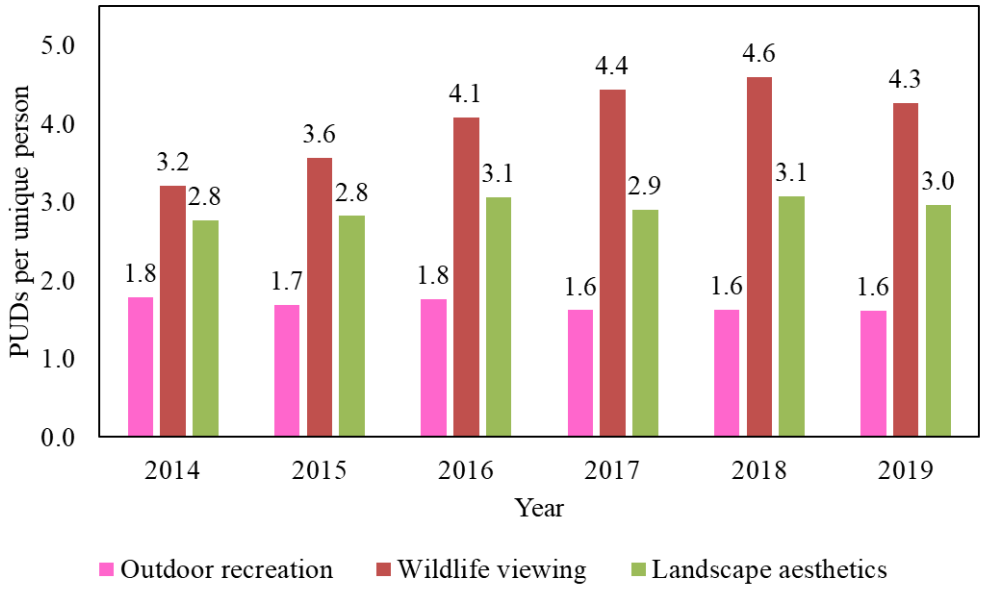

Figure S.2. Annual PUDs per user from 2014 to 2019.