Learning to Drive Anywhere via Regional Channel Attention

Anonymous Author(s)
Affiliation
Address
email

Abstract: Human drivers can seamlessly adapt their driving decisions across geographical locations with diverse conditions and rules of the road, e.g., left vs. right-hand traffic. In contrast, existing models for autonomous driving have been thus far only deployed within restricted operational domains, i.e., without accounting for varying driving behaviors across locations or model scalability. In this work, we propose GeCo, a single geographically-aware conditional imitation learning (CIL) model that can efficiently learn from heterogeneous and globally distributed data with dynamic environmental, traffic, and social characteristics. Our key insight is to introduce a high-capacity, geo-location-based channel attention mechanism that effectively adapts to local nuances while also flexibly modeling similarities among regions in a data-driven manner. By optimizing a contrastive imitation objective, our proposed approach can efficiently scale across the inherently imbalanced data distributions and location-dependent events. We demonstrate the benefits of our GeCo agent across multiple datasets, cities, and scalable deployment paradigms, i.e., centralized, semi-supervised, and distributed agent training. Specifically, GeCo outperforms CIL baselines by over 14% in open-loop evaluation and 30% in closed-loop testing on CARLA.

Keywords: Global-scale Autonomous Driving, Imitation Learning, Transformer

1 Introduction

Driving at scale involves a complex and nuanced decision-making process across diverse social and environmental conditions. For instance, a driving model for turning at intersections in San Francisco, CA may be able to generalize relatively well when deployed in Washington, DC, over 2400 miles away. However, when deployed just north of it in New York City, the model would discover that right turns on red, which have been banned in the city [1], could result in social disruption and potentially unsafe conditions at intersections. When deployed at intersections in Pittsburgh, PA, the model may begin accelerating at a green light only to be confounded by the frequent occurrence of the Pittsburgh left [2], resulting in frequent and uncomfortable braking. These examples illustrate how the lack of modeling of location-based traffic behavior and social norms can lead to potentially safety-critical consequences. Beyond city-level variability, failing to accurately account for state or country-level differences in traffic regulations and social norms can also have dire consequences, e.g., from the directionality of travel [3] to varying maximal speed limitations [4] or yielding expectations [5]. How can we design and learn models that flexibly accommodate the heterogeneous data encountered across challenging and diverse geographical, environmental, and social conditions?

Despite recent advances in decision-making models for autonomous driving, models are often trained and evaluated within limited operational domains, i.e., a handful of geographical regions and social conditions (e.g., Waymo's service in Pheonix, AZ, and San Francisco, CA [6]). Autonomous driving benchmarks are often collected in a handful of cities and routes [7–10]. Existing frameworks for learning to drive (e.g., [11–15]) train a single policy without considering geo-location or policy

Submitted to the 7th Conference on Robot Learning (CoRL 2023). Do not distribute.

adaptation. While such methods may be used to train and adapt different regional models, there are often many similarities which can be shared among the different locations and benefit potentially small and rare datasets with imbalanced distributions. In this work, we propose an approach for learning to modulate predictions across settings and locations, i.e., *even in seemingly similar visual settings*, within a single driving agent.

While prior works have leveraged transfer learning [16–22], e.g., through access to unlabeled data of a target domain, or introduced internal layers that learn to adapt model output across various domains [23–27, 16, 28–30], these methods have exclusively focused on low-level object classification and detection tasks, and have not explicitly accounted for geographical priors or reasoning. In contrast, we study end-to-end models for learning safe perception and decision-making in intricate 3D navigation scenarios. In this case, to avoid a potential accident, perception and action characteristics must both be carefully tuned to consider geographical location when reasoning over traffic maneuvers and predicting social behavior. Moreover, the training process of our sensorimotor models may require order-of-magnitude higher sample complexity, i.e., due to the higher rarity of policy-level events and intricate maneuvers [31]. Thus, geo-aware model capacity should be explored jointly with approaches for efficient adaptation and parameter sharing, as we do in this work.

Contributions: We make **three key contributions** towards autonomous systems at scale: 1) We revisit current end-to-end driving models to identify limitations in learning from heterogeneous and distributed data sources. In particular, we build on recent advances in transformer-based models [32, 33] for learning high-capacity, geo-aware imitation learning agents that can adapt across geographical locations while sharing parameters and computation within a single network. 2) To facilitate efficient training across inherently imbalanced data distributions and maneuvers, we further generalize conditional imitation learning by designing a *supervised contrastive loss* over conditional commands and locations. 3) We combine three public autonomous driving datasets collected by different companies and platforms across 11 locations to extensively evaluate the impact of the proposed scalable learning framework. To understand generalization across diverse use-cases and model training regimens, we comprehensively analyze the benefits of our framework for various scalable deployment scenarios, including centralized (i.e., within a single company or server with shared raw data logs), distributed (i.e., with scalable federated computation), and semi-supervised (i.e., with unlabeled data) training.

2 Related Work

Learning to Drive from Demonstrations: Despite impressive recent advances in learning to drive, approaches often leverage simple navigation tasks, i.e., lane following, intersection turning, and basic collision avoidance (e.g., with CIL [12, 34, 15, 35–39, 13]), or short real-world routes in a handful of locations (e.g., [40, 11, 41, 42, 14, 43, 10, 44–46, 7]). We note that GPS localization in prior approaches may only be used to determine a next *high-level command at an intersection* [34], and not to learn regionally or socially appropriate decisions. Yet, training models among locations without such geo-awareness results in an ill-posed problem with ambiguous samples. Thus, our work can be seen as a natural generalization of goal-conditional imitation learning frameworks [34, 12, 13] to incorporate geographical information for learning a high-capacity and controllable model.

Domain Adaptation: Model adaptation, i.e., from a source to a target domain with unlabeled data, has mostly focused on segmentation and detection tasks [16–22]. However, the robustness and reliability of current domain adaptation techniques at large scale have been repetitively questioned [47–49]. Moreover, the aforementioned techniques have not been previously studied within the more complex end-to-end training paradigm for decision-making models. Particularly relevant to our study are approaches that learn universal object detection models [23, 25] via self-attention and weighing feature channels based on the output of multiple parallel layers (i.e., adapters [50]). In contrast, our proposed *cross-attention-based* network architecture can more effectively fuse geographically-derived and visual features while also outperforming adapter-based methods [23].

Benefits of Contrastive Learning: Researchers have been increasingly exploring the benefits of contrastive learning frameworks for learning generalized representations, even under imbalanced or long-tail settings [51–55]. Our main use-case inherently involves learning over diverse and imbalanced underlying data distributions. For instance, a Tesla may suddenly trigger a warning in a challenging scenario or an unsupported region, in which case small amounts of demonstration data from the driver may be collected and available for training. Moreover, although diverse and rare traffic scenarios can occur within any local city region or country, the *underlying distribution of such events* can significantly shift among locations. Recently, Mandi et al. demonstrated the benefits of *unsupervised contrastive learning* for improved imitation learning within simple robotic use-cases [56]. Instead, we demonstrate the benefits of *supervised contrastive learning* techniques (e.g., [54]) by designing a novel loss function for conditional imitation learning frameworks at scale.

Distributed Learning to Drive: We comprehensively analyze our proposed approach across training paradigms suitable for scalable deployment in order to ensure the generalization of our findings. In particular, Federated Learning (FL) provides a natural framework for implementing GeCo in the real-world. The goal of FL to drive is to train a global model leveraging distributed data and models from different agents [57], i.e., where agents may avoid sharing raw driving logs and data due to various privacy and efficiency considerations. However, dealing with data heterogeneity among agents [58–60] remains a challenge. We demonstrate our novel geo-conditional mechanism to complement current federated learning algorithms. Somewhat surprisingly, our FL model variants result in outperformance compared to the centralized-trained counterparts due to the effective regional bias handling. We note that this is without having to share potentially sensitive geographical information, as our embedding matrix (defined in Sec. 3) is kept local and private in our implementation.

3 Method

We propose a geo-conditional (GeCo) agent which generalizes existing conditional imitation learning methods [34, 13] through two key aspects. First, we propose a novel network structure that leverages a *multi-head transformer module* for geo-aware adaptation of visual features across regions (Sec. 3.2). Second, we design a *contrastive learning objective* which regularizes training and addresses imbalances across locations and capture settings (Sec. 3.3). An overview of our approach is depicted in Fig. 1.

3.1 Problem Definition

Our objective is to learn a goal-directed agent that can effectively reason over varying traffic rules and social norms in complex and dynamic real-world settings. We leverage offline approaches relying on learning from driver demonstrations [61, 62, 40, 63, 13] as they can safely learn to map sensor observations to actions, i.e., as opposed to interactive methods [64, 65, 35, 66, 67]. As an example use-case, consider a deployed Tesla or Waymo fleet encountering challenging settings beyond its current constrained and geo-fenced deployment [68, 69]. Here, a human can take-over and demonstrate desired driving behavior which can subsequently be uploaded to a shared cloud server (i.e., centralized training) or updated to improve the model locally (i.e., federated training, we consider both cases in Sec. 3.4). However, current end-to-end agents that learn to drive in a data-driven manner, e.g., based on CIL [34, 12, 70, 13], do not differentiate among regional norms.

Geo-Conditional Imitation Learning: We assume a dataset of demonstrations $\mathcal{D} = \{(\mathbf{x}, \mathbf{y})\}_{i=1}^{N}$, i.e., measurements of $\mathbf{x} = (\mathbf{I}, c, v, \mathbf{g}) \in \mathcal{X}$, where $\mathbf{I} \in \mathbb{R}^{W \times H \times 3}$ is an image of the current environment, $v \in \mathbb{R}$ is the speed, $c \in \mathbb{N}$ is a navigation command [34, 12], $\mathbf{g} \in \{0, 1\}^{G}$ is a region index encoded as a one-hot vector over a total G regions, and corresponding action labels $\mathbf{y} \in \mathcal{Y}$ based on human drivers. Consistently with prior work [71–73, 36, 74, 75], we predict a waypoint-based label in the bird's eye view over the next five planned locations (2.5 seconds), such that $\mathbf{y} = \{\mathbf{w}_t\}_{t=1}^{5}$ and $\mathbf{w}_i \in \mathcal{R}^2$. The high-level waypoint output in the bird's eye view can also help standardize policy decisions across globally distributed platforms with heterogeneous sensor

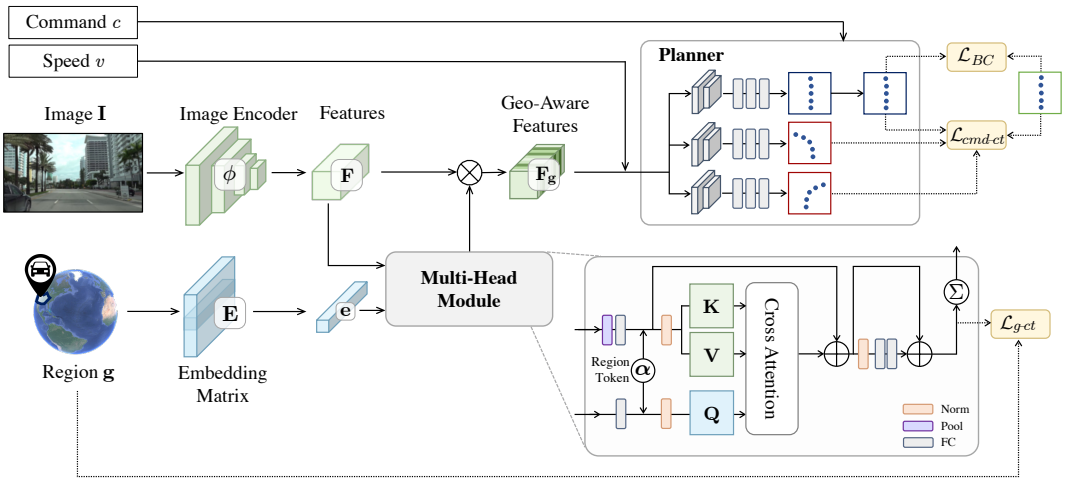

Figure 1: **Model Overview.** Our model maps image, region, speed, and conditional command observations to future decisions, parameterized as waypoints in the map view. To efficiently learn a high-capacity model, we leverage a multi-head cross attention module which fuses and adapts internal representations in a geo-aware manner. Our imitation objective, defined over human-demonstrated waypoints (outlined in green), the other command branches (outlined in red), and the predicted weights by the multi-head module, regularizes model optimization under diverse data distributions.

configurations [41]. In our work, we experiment with various definitions for $\mathbf{g}$ (manually defined city labels and unsupervised neighborhood-level labels, these do not require precise GPS localization). Moreover, we note GeCo does not rely on image-level perception labels or high-definition map information. As such, it can be trained based on cheaply collected GPS-based waypoint labels and benefit from rapid advancements in positioning technology (analysis of localization noise when training GeCo models can be found in the supplementary). We train a geographically-aware policy function $\pi : \mathcal{X} \rightarrow \mathcal{Y}$ using supervised learning [12, 37, 13]. Learning the policy π in geo-conditional imitation learning requires carefully fusing image and geo-location information, i.e., as opposed to just basic concatenation. Next, we introduce our network architecture which efficiently generalizes the branch-based architectures of CIL-based approaches [34].

3.2 Geo-Conditional Transformer Module

To learn a scalable policy function, i.e., across cities, countries, and platforms, we design a single network which adapts its decisions based on an efficient multi-head module. The mechanism is motivated by transformer [33, 32, 76], with three main aspects. First, the queries are only conditioned on the part of the input, i.e., the region-based features. Second, we do not use spatial attention as in ViT-type architectures [32], but instead learn a low-dimensional channel weight vector which can be trained efficiently [76]. Third, while multi-head mechanisms have been used by prior methods [32], we propose to jointly predict a scalar weight for each head prior to the summation of the heads. This formulation is analogous to a mixture or adapter-based model [77, 27]. Our domain attention mechanism is implemented via a *region token* that enables the model to specialize the heads to specific domains or tasks and subsequently combine the heads based on the current appropriate region and decision. For instance, we identify the emergence of traffic rules, such as left vs. right-hand driving when inspecting the output of the learned heads in Sec. 4.

Our model first extracts image features from the input image $\mathbf{I}$. Subsequently, a multi-head transformer module computes channel weights for the features based on the current region definition $\mathbf{g}$. Finally, the planner utilizes the re-weighted geo-aware features $\mathbf{F_g}$ and speed information v to generate waypoints for different commands. The command input c then selects the required waypoints $\hat{\mathbf{y}}$ for execution. The multi-head transformer module takes as input *visual features* extracted using a ResNet-34 encoder ϕ [78], $\mathbf{F} = \phi(\mathbf{I}) \in \mathbb{R}^{8\times13\times C}$ and a *regional embedding* $\mathbf{e} = \mathbf{g}^\top\mathbf{E}$

(assuming a column vector $\mathbf{g}$), extracted from a trainable embedding matrix $\mathbf{E} \in \mathbb{R}^{G \times C}$ [79]. We use $C = 512$ such that the output of the multi-head module is a 512-dimensional vector for weighting each channel in $\mathbf{F}$ and computing the geo-aware features $\mathbf{F_g}$ using the weighted and summed H output heads (Eqn. 3). The visual features are pooled (to accommodate the channel-wise attention), processed through a Fully Connected (FC) layer, and concatenated with a *region token* $\boldsymbol{\alpha}$, $\mathbf{z_I} = [\boldsymbol{\alpha}, \text{FC}((\text{Pool}(\mathbf{F}))] \in \mathbb{R}^{(C+1) \times d}$, where $d = 128$ sets the number of hidden units. $\boldsymbol{\alpha}$ will be updated and used to weigh the multiple heads at the output of the module, as shown in Eqn. 3. Similarly, the region embedding $\mathbf{z_g} = [\boldsymbol{\alpha},\ \text{FC}(\mathbf{e})] \in \mathbb{R}^{(C+1) \times d}$. The computation steps for the geo-aware transformer can then be summarized as:

$$\mathbf{z} = \mathbf{z_I} + \text{Attention}(\text{LN}(\mathbf{z_I}), \text{LN}(\mathbf{z_g})) \tag{1}$$

$$\hat{\mathbf{z}} = \mathbf{z} \ + \text{MLP}(\text{LN}(\mathbf{z})) \tag{2}$$

$$\mathbf{F_g} = \sum_{h=1}^{H} (\hat{\boldsymbol{\alpha}}_h \hat{\mathbf{z}}_{h,2:C+1}) \otimes \mathbf{F} \tag{3}$$

where LN denotes Layer Normalization, $\otimes$ denotes channel-wise multiplication, and $\hat{\boldsymbol{\alpha}}$ is the updated region token values. $\mathbf{z}$ in the second step is pooled before addition to make the shape consistent. We follow ViT [32] to compute attention as

$$\text{Attention}(\mathbf{z_I}, \mathbf{z_g}) = \text{softmax}\Big(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d}}\Big)\mathbf{V} \tag{4}$$

where $\mathbf{Q} = \mathbf{z_g}\mathbf{W}^Q$, $\mathbf{K} = \mathbf{z_I}\mathbf{W}^K$, $\mathbf{V} = \mathbf{z_I}\mathbf{W}^V$ and $\mathbf{W}^Q \in \mathbb{R}^{d \times d}$, $\mathbf{W}^K \in \mathbb{R}^{d \times d}$ and $\mathbf{W}^V \in \mathbb{R}^{d \times d}$ are learned matrices. Unlike ViT, we do not merge the multiple heads by concatenating such that there are H outputs, each C+1-dimensional, i.e., $\hat{\mathbf{z}} \in \mathbb{R}^{H \times (C+1)}$. Here, the weights for the h-th head are stored at the first index of the head output vector, i.e., $\hat{\boldsymbol{\alpha}}_h = \hat{\mathbf{z}}_{h,1}$. The adapted geo-aware features are then given to a command-conditional branch as shown in Fig. 1 for predicting the final waypoints. To optimize the network, we leverage a contrastive loss function over maneuvers and regional decisions, as discussed next.

3.3 Contrastive Imitation Learning

Loss Function: Standard supervised learning approaches for imitation learning leverage an L_1 loss, i.e., behavior cloning, between predicted and demonstrated ground-truth waypoints [67, 12]. However, in our case of highly heterogeneous and imbalanced data over maneuvers and regions, this loss can result in poor performance and overfitting to local biases [80, 54]. In particular, the distribution of both conditional commands c and regions $\mathbf{g}$ can be highly skewed, with certain critical events (e.g., turns) occurring at a much lower frequency. While we employ a branched architecture [34, 13], a subset of the branches may be trained over a fraction of the total samples, i.e., with most updating the 'forward' branch. We hypothesize that such imbalances can introduce noisy predictions from poorly-trained branches. When adding the additional complexity of learning region-conditional policies, issues in data imbalance and heterogeneity compound. To tackle this practical safety-critical issue, Chen et al. [62] employed a privileged teacher (i.e., learned from complete ground truth observations of the 3D surroundings instead of raw images) that can be used for additional sampling and data augmentation. However, training such a privileged expert requires extensive annotation of real-world data, which is not scalable. Instead, we propose to introduce command and region-contrastive objectives as a simple and effective strategy for improving model optimization and providing more supervision when handling imbalanced data. In our analysis, we demonstrate the utility of this approach for both vanilla CIL and the proposed geo-CIL. As far as we are aware, we are the first to empirically analyze such benefits for imitation-learned driving agents at scale.

We propose to incorporate two additional terms in addition to the main behavior cloning loss, $\mathcal{L}_{BC}$. The total loss can be computed as

$$\mathcal{L} = \mathcal{L}_{BC} + \lambda_c \mathcal{L}_{cmd-ct} + \lambda_g \mathcal{L}_{g-ct} \tag{5}$$

where λ_c, λ_g are hyperparameters, $\mathcal{L}_{cmd-ct}$ and $\mathcal{L}_{g-ct}$ are contrastive losses. Next, we define each of the proposed loss terms.

Command Contrastive Loss: The branched conditional command architecture of CIL updates each branch based on a subset of the data samples in each batch $\mathcal{B}$. As some commands are highly underrepresented in natural driving data, we propose a command contrastive loss that leverages *predictions for other commands for the same sample* as negative examples,

$$\mathcal{L}_{cmd-ct} = -\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\log\frac{\exp(d(\hat{\mathbf{y}}_i^{c_i}, \mathbf{y}_i)/\tau)}{\sum_c \exp(d(\hat{\mathbf{y}}_i^{c}, \mathbf{y}_i)/\tau)} \tag{6}$$

where the loss is computed over a single positive example with prediction outputted by the ground truth command branch c_i, and the other predictions (i.e., hypothetical commands) as negatives. d is a similarity function (we use negative L_2 distance) and $\tau \in \mathbb{R}^+$ is a scalar temperature parameter.

The command contrastive loss is a natural extension supervised contrastive learning [54] to our case of conditional imitation learning, with two key differences. First, we do not apply it to the feature space (as commonly done) but instead to the *output space* of the network, i.e., in order to better model differences among maneuvers. Second, the command contrastive loss is not computed over all other samples in the same batch with different commands as in standard supervised contrastive loss [54]. While this practice may work for simple classification tasks, in our case such other samples tend to also involve different driving situations. We found leveraging other samples in this manner when training an imitation model to degrade policy performance, most likely due to the added learning complexity and ambiguity. A similar reasoning can be applied to improve optimization of the geo-conditioned transformer module, as discussed next.

Geo-Contrastive Loss: While driving behavior across cities and regions can often be similar, in practice, the cities in our employed datasets (detailed in Sec. 4) all have unique local characteristics effectively modeled. Thus, we also propose to incorporate a region (i.e., city)-based contrastive loss. We apply the loss within the transformer module over different output head weights $\hat{\boldsymbol{\alpha}} \in \mathbb{R}^H$. Here, we follow standard contrastive loss implementation [54] and select the i-th sample as an anchor. During training, for the i-th sample in a batch, positive samples $\mathcal{P}(i)$ are defined within the same city, while negative samples $\mathcal{N}(i)$ from differing cities,

$$\mathcal{L}_{g-ct} = -\frac{1}{|\mathcal{B}|}\sum_{i\in\mathcal{B}}\frac{1}{|\mathcal{P}(i)|}\log\frac{\sum_{p\in\mathcal{P}(i)}\exp(d(\hat{\boldsymbol{\alpha}}_i, \hat{\boldsymbol{\alpha}}_p)/\tau)}{\sum_{a\in\mathcal{A}(i)}\exp(d(\hat{\boldsymbol{\alpha}}_i, \hat{\boldsymbol{\alpha}}_a)/\tau)} \tag{7}$$

where $\mathcal{A}(i) \equiv \mathcal{P}(i) \cup \mathcal{N}(i)$ and d is a similarity function (we use negative L_2 distance).

3.4 Scalable Training Settings

We comprehensively analyze the training of our model using three different scalable deployment settings. First, in ***Centralized Learning (CL)***, agents are able to share all sensor data and geographical information with a centralized server. Consequently, the server conducts supervised learning of our GeCo model over all of the raw data. To further analyze model scalability, we implement a ***Semi-Supervised Learning (SSL)*** model, which can leverage ample unlabeled data that may be available across locations (we follow [13]). Finally, we study the applicability of our findings within federated learning approaches, as sharing raw sensor data can be inefficient or even potentially undesirable. For instance, our GeCo agent may be distributed over numerous heterogeneous data sources with various constraints, i.e., local regulations, legal authorities, and privacy requirements or preferences. Thus, we also analyze a ***Federated Learning (FL)*** agent which does not require sharing raw and geographical information with a server. To fully understand the role of our proposed network structure within such paradigms, we optimize GeCo using two federated learning algorithms, FedAvg[57] and FedDyn[58]. We note that the geographical embedding matrix $\mathbf{E}$ which contains city-level information remains locally updated on each agent (i.e., akin to a form of local model personalization). In this manner, $\mathbf{E}$ reduces to a row vector as the embedding vector for the specific region (city in our

implementation). We note that this results in the removal of the geo-contrastive loss term $\mathcal{L}_{g-ct}$ in Eqn. 7. The supplementary contains additional details regarding our implementation.

4 Experiments

In this section, we first introduce our combined multi-city benchmark extracted from multiple publicly available datasets. Specifically, we present ablation studies for various model design choices and loss terms. To understand the benefits of the proposed framework on various training schemes, we also report analysis with three different training paradigms. This ensures our model and findings are relevant across real-world use-cases, e.g., with efficient distributed settings at large-scale.

4.1 Datasets and Metrics

To learn a global scale driving policy, we conduct training on data from three different datasets including Argoverse 2 (***AV2***) [9], nuScenes (***nS***) [10] and Waymo (***Waymo***) [7]. While these datasets do not have any official waypoints prediction benchmark, we extract these from the provided raw data logs. Specifically for each frame, we processing the raw data to get the future 2.5s as ground truth waypoints, current velocity, a front-view RGB image, navigational commands and a city-level information. The data spans 11 cities. We split the data into training, validation and testing data. Our split utilizes 190k, 20k, and 35k training samples for AV2, nS and Waymo datasets respectively. We follow standard evaluation using Average L_2 Displacement Error (ADE) and Final L_2 Displacement Error (FDE) over future waypoints in the BEV space. We also evaluate closed-loop policy performance using CARLA [81]. While CARLA benchmarks do not generally involve regional modeling, we simulate left-hand driving and town-varying behavior of agents. Our supplementary provides additional details and experiments, e.g., regarding unsupervised geo-location clustering mechanisms and closed-loop evaluation.

4.2 Results

Model and Loss Ablation: We first study the underlying architecture of the model in Table 1. We find that replacing intermediate image-level heatmaps (used in several CIL-based baselines [62, 13]) with fully-connected layers provides improved reasoning for our diverse perspective settings (reducing ADE from 1.24 to 1.16). We also demonstrate our geo-conditional transformer framework with three heads to outperform other supervision choices, e.g., embedding concatenation and supervision as an auxiliary prediction task as in Ayush et al. [82] (1.09 vs. 1.15 ADE). GeCo also outperforms another attention-based method e.g., Hybrid ViT [32] on concatenated image-city features (1.09 vs. 1.20 ADE), which validates its efficiency on adaptation. Moreover, the proposed geo-conditional module can be used to increase the modeling capacity of the agent, and thus can scale beyond simple task supervision. We also find a holistic effect among the proposed loss terms, with a combination leading to the best results (1.93 vs 2.08 FDE for the vanilla behavior cloning loss). Our qualitative results show GeCo to better handle diverse traffic regulations and social norms, e.g., turning right (wider turn) in Singapore and yielding a 'Pittsburgh left' vehicle in Pittsburgh (Fig. 2).

Table 1: **Ablative Studies on Model Architecture, Geo-Conditional Module and Loss.** We start from CIL architectures [62, 13] and gradually add different components losses to get GeCo.

Ablation	Method	ADE	FDE
CIL Architecture	CIL [62]	1.32	2.55
	BEV Planner [13]	1.24	2.45
	Our Planner	**1.16**	**2.17**
Geo-CIL Architecture	Concatenation	1.14	2.12
	Task Supervision [82]	1.15	2.24
	Hybrid ViT [32]	1.20	2.30
	Universal Adapter [83]	1.10	2.11
	Geo Transformer w/ $\mathcal{L}_{BC}$	**1.09**	**2.08**
Loss Function	$\mathcal{L}_{BC}, \mathcal{L}_{cmd-ct}$	1.07	1.97
	$\mathcal{L}_{BC}, \mathcal{L}_{g-ct}$	1.06	2.00
	GeCo ($\mathcal{L}$)	**1.05**	**1.93**

Training Paradigms: Table 2 reports the impact of various model training schemes on ADE performance (additional details, including FDE-based analysis, are in the supplementary). We observe consistent improvements across paradigms and cities even *with severe data imbalance*. Moreover, leveraging unlabeled YouTube data for each city results in further gains, specifically for cities with lesser data (MTV, PAO, SGP, and PHX). For instance, MTV improves from 1.40 to 1.23 ADE due to

Table 2: **Evaluating GeCo with Different Training Paradigms.** GeCo efficiently integrates into various training paradigms (CL-Centralized Learning, SSL-Semi-Supervised Learning, and FL-Federated Learning). ADE is computed across the 11 cities in our dataset. Our planner is our proposed architecture for direct image-to-BEV prediction (without the geolocation information or introduced auxiliary loss terms, see supplementary for additional architecture details).

Settings	**Method**	**Avg**	*PIT*	*WDC*	*MIA*	*ATX*	*PAO*	*DTW*	*BOS*	*SGP*	*PHX*	*SFO*	*MTV*
CL	CIL [62]	1.32	1.14	1.40	1.47	1.23	1.49	1.01	0.89	1.09	1.65	1.67	1.48
	CILRS [12]	1.27	1.18	1.28	1.43	1.16	1.52	1.11	0.84	1.02	1.39	1.60	1.40
	BEV Planner [13]	1.24	1.18	1.01	1.34	1.23	1.55	1.03	0.90	1.07	1.38	1.58	**1.39**
	TCP [38]	1.22	1.09	1.23	1.41	1.14	1.47	0.99	0.87	1.01	1.39	1.49	1.40
	Our Planner	1.16	1.24	1.12	1.12	1.38	1.39	1.02	0.92	1.10	1.08	0.89	1.41
	GeCo	**1.05**	**1.12**	**0.96**	**0.95**	**1.16**	**1.31**	**0.89**	**0.82**	**1.03**	**0.98**	**0.83**	1.40
SSL	SelfD [13]	1.02	1.13	1.03	1.01	1.25	1.26	0.93	0.80	0.95	**0.82**	0.79	1.29
	GeCo	**0.97**	**1.06**	**0.93**	**0.92**	**1.19**	**1.24**	**0.89**	**0.76**	**0.94**	0.84	**0.75**	**1.23**
FL	FedAvg [57]	1.42	1.38	1.43	1.41	1.73	1.63	1.23	0.93	1.21	1.53	1.42	1.64
	FedDyn [58]	1.19	1.23	1.15	1.21	1.59	**1.51**	1.11	0.797	0.95	0.97	0.99	1.30
	GeCo (FedAvg)	1.20	1.23	1.06	1.04	1.62	1.61	1.00	0.81	1.07	1.33	1.01	1.41
	GeCo (FedDyn)	**0.98**	**1.08**	**0.91**	**0.91**	**1.50**	1.54	**0.90**	**0.68**	**0.82**	**0.62**	**0.70**	**1.12**

the unlabeled data, showing the importance of this mechanism for our use-case. Overall, the model is shown to outperform the baseline of Zhang et al. [13], which does not leverage geo-location information. As expected, federated learning algorithms under-perform their centralized counterparts in Table 2. In contrast, we find federated learning with GeCo to surpass centralized training, e.g., from 1.05 to 0.98, potentially due to better handling of local biases. Additional ablations and experiments showing the complexity of our real-world modeling task can be found in the supplementary.

5 Conclusion

We envision large-scale navigation agents that can seamlessly operate in heterogeneous and distributed locations. Towards this goal, our work introduces an efficient framework for training and adapting a universal high-capacity navigation agent across diverse locations and settings. Using our proposed agent, fleets of vehicles can increasingly grow their operation capacity to novel conditions, i.e., by involving humans and collecting both unlabeled or labeled demonstration data for policy training. Nonetheless, effectively incorporating geo-awareness into driving models remains a challenging under-explored research problem, as discussed next.

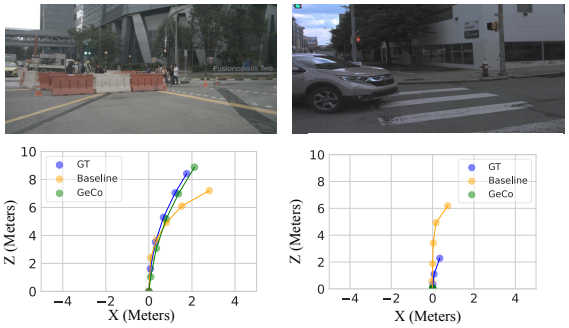

Figure 2: **Qualitative Results.** We plot predicted waypoints in the BEV for comparison. GeCo exhibits robustness in region-specific cases, including turning right (wider turn) in Singapore, yielding a 'Pittsburgh left' vehicle in Pittsburgh.

6 Limitations

Despite the multiple publicly available datasets used in our experiments, the diversity in existing benchmarks is still limited, i.e., compared to the vast diversity of geo-locations and events that an agent may encounter in the real-world. Besides Singapore, which provides a challenging generalization use-case , data logs in current datasets are often captured over short drives and are biased towards the US. Thus, our framework requires further validation with larger-scale settings with increased diversity in the future. Here, while our approach for learning a unified model is motivated by human drivers that efficiently learn to adapt generalized skills across locations (including traffic direction), it can be potentially challenging to learn a single model across drastically differing locations. Finally, incorporating various explicit constraints and specifications (e.g., of local traffic rules) could also be studied in the future in order to enable efficient agent adaptation.

References

[1] Wikipedia. Turn on red. *https://en.wikipedia.org/wiki/Turn_on_red*, 2022.

[2] Wikipedia. Pittsburgh left. *https://en.wikipedia.org/wiki/Pittsburgh_left*, 2019.

[3] ChartsBin. Worldwide driving orientation by country. *http://chartsbin.com/view/edr*, 2009.

[4] VividMaps. What traffic rules are different in different countries. *https://vividmaps.com/what-traffic-rules-are-different-in-different-countries/*, 2019.

[5] Wikipedia. Priority to the right. *https://en.wikipedia.org/wiki/Priority_to_the_right*, 2022.

[6] Waymo one. *https://waymo.com/waymo-one/*, 2022.

[7] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, 2020.

[8] X. Huang, X. Cheng, Q. Geng, B. Cao, D. Zhou, P. Wang, Y. Lin, and R. Yang. The apolloscape dataset for autonomous driving. In *CVPRW*, 2018.

[9] B. Wilson, W. Qi, T. Agarwal, J. Lambert, J. Singh, S. Khandelwal, B. Pan, R. Kumar, A. Hartnett, J. K. Pontes, et al. Argoverse 2.0: Next generation datasets for self-driving perception and forecasting. In *NeurIPS*, 2021.

[10] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020.

[11] M. Bansal, A. Krizhevsky, and A. Ogale. ChauffeurNet: Learning to drive by imitating the best and synthesizing the worst. In *RSS*, 2019.

[12] F. Codevilla, E. Santana, A. M. López, and A. Gaidon. Exploring the limitations of behavior cloning for autonomous driving. In *ICCV*, 2019.

[13] J. Zhang, R. Zhu, and E. Ohn-Bar. SelfD: Self-learning large-scale driving policies from the web. In *CVPR*, 2022.

[14] M. Bojarski, D. Del Testa, D. Dworakowski, B. Firner, B. Flepp, P. Goyal, L. D. Jackel, M. Monfort, U. Muller, J. Zhang, et al. End to end learning for self-driving cars. In *arXiv*, 2016.

[15] D. Chen, V. Koltun, and P. Krähenbühl. Learning to drive from a world on rails. In *ICCV*, 2021.

[16] N. Dvornik, C. Schmid, and J. Mairal. Selecting relevant features from a multi-domain representation for few-shot classification. In *ECCV*, 2020.

[17] Y. Wang, J. Peng, and Z. Zhang. Uncertainty-aware pseudo label refinery for domain adaptive semantic segmentation. In *ICCV*, 2021.

[18] K. Saito, Y. Ushiku, and T. Harada. Asymmetric tri-training for unsupervised domain adaptation. In *ICML*, 2017.

[19] J. Yang, S. Shi, Z. Wang, H. Li, and X. Qi. ST3D: Self-training for unsupervised domain adaptation on 3d object detection. In *CVPR*, 2021.

[20] Z. Lin, D. Ramanan, and A. Bansal. Streaming self-training via domain-agnostic unlabeled images. In *arXiv*, 2021.

[21] Y. Liu, W. Zhang, and J. Wang. Source-free domain adaptation for semantic segmentation. In *CVPR*, 2021.

[22] K. Wang, X. Gao, Y. Zhao, X. Li, D. Dou, and C.-Z. Xu. Pay attention to features, transfer learn faster cnns. In *ICLR*, 2019.

[23] X. Wang, Z. Cai, D. Gao, and N. Vasconcelos. Towards universal object detection by domain attention. In *CVPR*, 2019.

[24] I. Kokkinos. Ubernet: Training a universal convolutional neural network for low-, mid-, and high-level vision using diverse datasets and limited memory. In *CVPR*, 2017.

[25] W.-H. Li, X. Liu, and H. Bilen. Cross-domain few-shot learning with task-specific adapters. In *CVPR*, 2022.

[26] S.-A. Rebuffi, H. Bilen, and A. Vedaldi. Learning multiple visual domains with residual adapters. In *NeurIPS*, 2017.

[27] W.-H. Li, X. Liu, and H. Bilen. Universal representation learning from multiple domains for few-shot classification. In *ICCV*, 2021.

[28] E. Perez, F. Strub, H. De Vries, V. Dumoulin, and A. Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI*, 2018.

[29] P. Bateni, R. Goyal, V. Masrani, F. Wood, and L. Sigal. Improved few-shot visual classification. In *CVPR*, 2020.

[30] J. Requeima, J. Gordon, J. Bronskill, S. Nowozin, and R. E. Turner. Fast and flexible multi-task classification using conditional neural adaptive processes. In *NeurIPS*, 2019.

[31] S. Shalev-Shwartz and A. Shashua. On the sample complexity of end-to-end training vs. semantic abstraction training. In *arXiv*, 2016.

[32] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[34] F. Codevilla, M. Miiller, A. López, V. Koltun, and A. Dosovitskiy. End-to-end driving via conditional imitation learning. In *ICRA*, 2018.

[35] E. Ohn-Bar, A. Prakash, A. Behl, K. Chitta, and A. Geiger. Learning situational driving. In *CVPR*, 2020.

[36] N. Hanselmann, K. Renz, K. Chitta, A. Bhattacharyya, and A. Geiger. King: Generating safety-critical driving scenarios for robust imitation via kinematics gradients. In *ECCV*, 2022.

[37] K. Chitta, A. Prakash, B. Jaeger, Z. Yu, K. Renz, and A. Geiger. Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. *PAMI*, 2022.

[38] P. Wu, X. Jia, L. Chen, J. Yan, H. Li, and Y. Qiao. Trajectory-guided control prediction for end-to-end autonomous driving: A simple yet strong baseline. In *NeurIPS*, 2022.

[39] A. Hu, G. Corrado, N. Griffiths, Z. Murez, C. Gurau, H. Yeo, A. Kendall, R. Cipolla, and J. Shotton. Model-based imitation learning for urban driving. *NeurIPS*, 2022.

[40] J. Hawke, R. Shen, C. Gurau, S. Sharma, D. Reda, N. Nikolov, P. Mazur, S. Micklethwaite, N. Griffiths, A. Shah, et al. Urban driving with conditional imitation learning. In *ICRA*, 2020.

[41] M. Müller, A. Dosovitskiy, B. Ghanem, and V. Koltun. Driving policy transfer via modularity and abstraction. In *CoRL*, 2018.

[42] W. Zeng, W. Luo, S. Suo, A. Sadat, B. Yang, S. Casas, and R. Urtasun. End-to-end interpretable neural motion planner. In *CVPR*, 2019.

[43] L. L. Li, B. Yang, M. Liang, W. Zeng, M. Ren, S. Segal, and R. Urtasun. End-to-end contextual perception and prediction with interaction transformer. In *IROS*, 2020.

[44] M.-F. Chang, J. Lambert, P. Sangkloy, J. Singh, S. Bak, A. Hartnett, D. Wang, P. Carr, S. Lucey, D. Ramanan, et al. Argoverse: 3d tracking and forecasting with rich maps. In *CVPR*, 2019.

[45] J. Houston, G. Zuidhof, L. Bergamini, Y. Ye, L. Chen, A. Jain, S. Omari, V. Iglovikov, and P. Ondruska. One thousand and one hours: Self-driving motion prediction dataset. In *CoRL*, 2021.

[46] X. Huang, P. Wang, X. Cheng, D. Zhou, Q. Geng, and R. Yang. The ApolloScape open dataset for autonomous driving and its application. In *IEEE PAMI*, 2020.

[47] V. Prabhu, R. R. Selvaraju, J. Hoffman, and N. Naik. Can domain adaptation make object recognition work for everyone? In *CVPR*, 2022.

[48] S. Sagawa, P. W. Koh, T. Lee, I. Gao, S. M. Xie, K. Shen, A. Kumar, W. Hu, M. Yasunaga, H. Marklund, et al. Extending the wilds benchmark for unsupervised adaptation. In *ICLR*, 2021.

[49] R. Taori, A. Dave, V. Shankar, N. Carlini, B. Recht, and L. Schmidt. Measuring robustness to natural distribution shifts in image classification. In *NeurIPS*, 2020.

[50] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. In *CVPR*, 2018.

[51] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.

[52] T. Bai, J. Chen, J. Zhao, B. Wen, X. Jiang, and A. Kot. Feature distillation with guided adversarial contrastive learning. In *arXiv*, 2020.

[53] J. Li, P. Zhou, C. Xiong, and S. C. Hoi. Prototypical contrastive learning of unsupervised representations. *ICLR*, 2021.

[54] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan. Supervised contrastive learning. In *NeurIPS*, 2020.

[55] J. Cui, Z. Zhong, S. Liu, B. Yu, and J. Jia. Parametric contrastive learning. In *ICCV*, 2021.

[56] Z. Mandi, F. Liu, K. Lee, and P. Abbeel. Towards more generalizable one-shot visual imitation learning. In *ICRA*, 2022.

[57] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas. Communication-efficient learning of deep networks from decentralized data. In *AISTATS*, 2017.

[58] D. A. E. Acar, Y. Zhao, R. M. Navarro, M. Mattina, P. N. Whatmough, and V. Saligrama. Federated learning based on dynamic regularization. In *ICLR*, 2021.

[59] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith. Federated optimization in heterogeneous networks. In *MLSys*, 2020.

[60] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh. Scaffold: Stochastic controlled averaging for federated learning. In *ICML*, 2020.

[61] T. Osa, J. Pajarinen, G. Neumann, J. A. Bagnell, P. Abbeel, J. Peters, et al. An algorithmic perspective on imitation learning. In *Foundations and Trends® in Robotics*, 2018.

[62] D. Chen, B. Zhou, V. Koltun, and P. Krähenbühl. Learning by cheating. In *CoRL*, 2020.

[63] O. Scheel, L. Bergamini, M. Wołczyk, B. Osiński, and P. Ondruska. Urban driver: Learning to drive from real-world demonstrations using policy gradients. In *CoRL*, 2021.

[64] X. Liang, T. Wang, L. Yang, and E. Xing. CIRL: Controllable imitative reinforcement learning for vision-based self-driving. In *ECCV*, 2018.

[65] Z. Zhang, A. Liniger, D. Dai, F. Yu, and L. Van Gool. End-to-end urban driving by imitating a reinforcement learning coach. In *ICCV*, 2021.

[66] S. Ross, G. Gordon, and D. Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *AISTATS*, 2011.

[67] A. Prakash, A. Behl, E. Ohn-Bar, K. Chitta, and A. Geiger. Exploring data aggregation in policy learning for vision-based urban autonomous driving. In *CVPR*, 2020.

[68] WIRED. Why people keep rear-ending self-driving cars. *https://www.wired.com/story/self-driving-car-crashes-rear-endings-why-charts-statistics/*, 2018.

[69] electrek. We tested tesla full self-driving beta in the blue ridge mountains, and it was scary. *https://electrek.co/2022/08/15/tesla-full-self-driving-beta-blue-ridge-mountains-scary/*, 2022.

[70] F. Codevilla, A. M. Lopez, V. Koltun, and A. Dosovitskiy. On offline evaluation of vision-based driving models. In *ECCV*, 2018.

[71] K. Mangalam, H. Girase, S. Agarwal, K.-H. Lee, E. Adeli, J. Malik, and A. Gaidon. It is not the journey but the destination: Endpoint conditioned trajectory prediction. In *ECCV*, 2020.

[72] N. Rhinehart, R. McAllister, and S. Levine. Deep imitative models for flexible inference, planning, and control. In *ICLR*, 2020.

[73] N. Rhinehart, K. Kitani, and P. Vernaza. R2p2: A reparameterized pushforward policy for diverse, precise generative path forecasting. In *ECCV*, 2018.

[74] K. Chitta, A. Prakash, B. Jaeger, Z. Yu, K. Renz, and A. Geiger. Transfuser: Imitation with transformer-based sensor fusion for autonomous driving. In *PAMI*, 2022.

[75] A. Prakash, K. Chitta, and A. Geiger. Multi-modal fusion transformer for end-to-end autonomous driving. In *CVPR*, 2021.

[76] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon. CBAM: Convolutional block attention module. In *ECCV*, 2018.

[77] X. Wang, F. Yu, L. Dunlap, Y.-A. Ma, R. Wang, A. Mirhoseini, T. Darrell, and J. E. Gonzalez. Deep mixture of experts via shallow embedding. In *ICML*, 2018.

[78] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[79] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv*, 2013.

[80] B. Kang, Y. Li, S. Xie, Z. Yuan, and J. Feng. Exploring balanced feature spaces for representation learning. In *ICLR*, 2020.

[81] A. Dosovitskiy, G. Ros, F. Codevilla, A. Lopez, and V. Koltun. CARLA: An open urban driving simulator. In *CoRL*, 2017.

[82] K. Ayush, B. Uzkent, C. Meng, K. Tanmay, M. Burke, D. Lobell, and S. Ermon. Geography-aware self-supervised learning. In *CVPR*, 2021.

[83] H. Bilen and A. Vedaldi. Universal representations: The missing link between faces, text, planktons, and cat breeds. In *arXiv*, 2017.