LongComp: Long-Tail Compositional Zero-Shot Generalization for Robust Trajectory Prediction

Benjamin Stoler[1,2] Jonathan Francis[1,3] Jean Oh[1]

***Abstract*— Methods for trajectory prediction in Autonomous Driving must contend with rare, safety-critical scenarios that make reliance on real-world data collection alone infeasible. To assess robustness under such conditions, we propose new long-tail evaluation settings that repartition datasets to create challenging out-of-distribution (OOD) test sets. We first introduce a safety-informed scenario factorization framework, which disentangles scenarios into discrete ego and social contexts. Building on analogies to compositional zero-shot image-labeling in Computer Vision, we then hold out novel context combinations to construct challenging closed-world and open-world settings. This process induces OOD performance gaps in future motion prediction of 5.0% and 14.7% in closed-world and open-world settings, respectively, relative to in-distribution performance for a state-of-the-art baseline. To improve generalization, we extend task-modular gating networks to operate within trajectory prediction models, and develop an auxiliary, difficulty-prediction head to refine internal representations. Our strategies jointly reduce the OOD performance gaps to 2.8% and 11.5% in the two settings, respectively, while still improving in-distribution performance.**

I. Introduction

In autonomous driving (AD), the long-tail distribution of safety-critical scenarios makes it infeasible to develop and assess systems by relying on large-scale, real-world data collection [1]. Understanding and improving model performance in out-of-distribution (OOD) deployment settings has thus become increasingly important, including in the essential motion prediction component [2], [3]. Recent work has turned to creating artificial distribution shifts to emulate such OOD deployment settings, repartitioning existing datasets into non-uniform train and test sets on the bases of e.g., safety-relevance scores [4] or cluster identities [5]. However, these bases are often highly entangled (i.e., single axes that conflate many underlying attributes), which results in interpretation and generalization challenges, as emphasized in the broader machine learning literature [6]–[8].

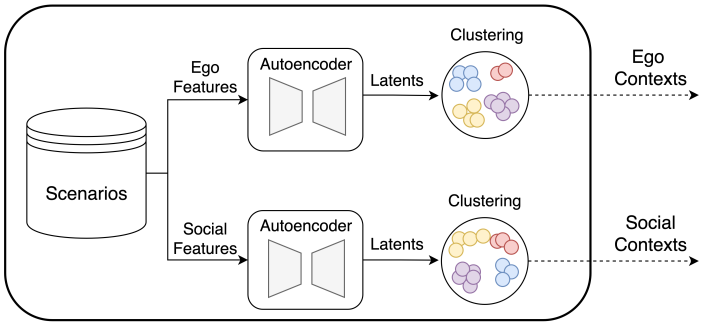
Recently, work in related machine learning tasks, such as labeling objects in images, has developed a relevant framework known as compositional zero-shot learning (CZSL) [9], [10]. In CZSL, object labels in images consist of pairs of object types and attributes. The zero-shot challenge is to train labeling models on some in-distribution (ID) *seen* subset of these pairs and test on both seen and OOD *unseen* pairs. These settings and resulting methodologies have demonstrated finer-grained evaluation of capabilities and improved model generalizability [10], [11]. It is thus appealing to extend the CZSL setting to AD tasks, especially given both the hierarchical manner in which humans operate vehicles [12], as well as the causal factorization of crash risks therein [13], [14]. In the context of driving, however, behavior generalization along safety-relevant, semantically meaningful axes has not yet been explored.

To address this gap, we propose and assess a safety-informed scenario factorization approach, enabling both the creation of challenging OOD evaluation settings and the development of robust, generalization methods for AD motion prediction. Existing studies on human driver error and risk assessment have organized contributing factors into pertinent categories: road infrastructure (road layout, signage quality), vehicle-related issues (mechanical reliability, maintenance), road user conditions (driver experience, mental state), behaviors of other road users (aggression, erraticism), and environmental conditions (lighting, weather) [13], [14]. Unfortunately, large motion prediction datasets typically lack much of this detailed information [15], [16].

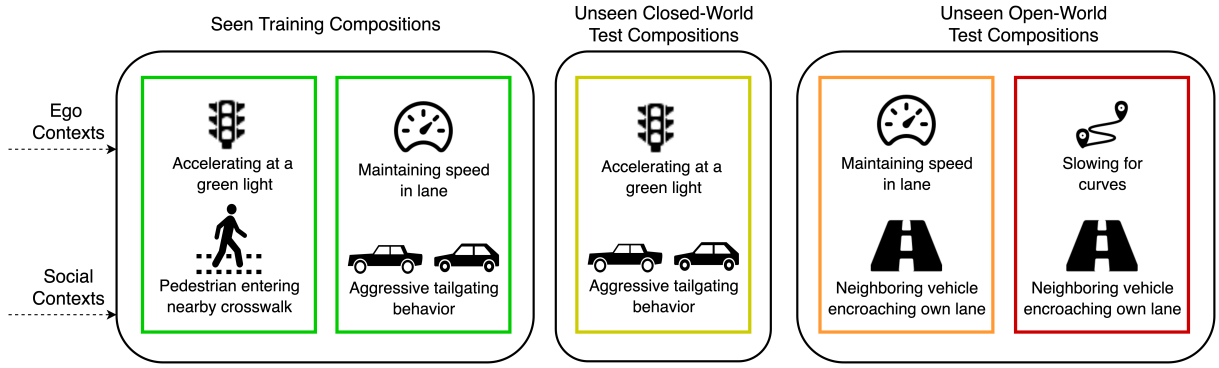
We thus derive two axes from available indicators: an "ego" context, capturing both *general* and *safety-relevant* factors (such as kinematics, map features, and deviation from the speed limit), and a "social" context, capturing relative kinematics as well as safety-criticality indicators (including closing speeds and minimum time-to-conflict point difference (mTTCP)), as shown in Figure 1a. We then create long-tail, zero-shot compositional closed-world and open-world settings on top of these paired contexts, with these axes serving as analogues to the object and attribute axes in the image-labeling CZSL setting. These settings entail splitting data into seen and unseen portions non-uniformly, holding out novel combinations of ego and social contexts to be tested on, as shown in Figure 1b. Note that, unlike the established image-labeling setting, our evaluation focuses on downstream task performance *within* these novel context combinations, rather than predicting a label for a scenario; to our knowledge, this is the **first** extension of CZSL concepts to trajectory prediction.

In these closed-world and open-world settings, we observe a significant OOD performance gap in trajectory prediction, when using WOMD [15] as an exemplar dataset and MTR [17] as a baseline prediction approach. To enhance OOD performance, we then develop and extend domain generalization techniques from the image-labeling CZSL setting. In particular, we adapt task modular gating networks [18] to operate directly in the bottleneck layer of baseline approaches and further enhance the intermediate representation with an auxiliary, difficulty-prediction head.

[1]School of Computer Science, Carnegie Mellon University, Email: {bstoler, jmf1, jeanoh}@cs.cmu.edu. [2]Stack AV Research Internship. [3]Bosch Center for Artificial Intelligence.



(a) Scenario factorization into paired ego and social contexts.



(b) Illustrative toy examples of context compositions.

Fig. 1: Overview of our framework. (a) Traffic scenarios are factorized and clustered along explicitly disentangled ego and social axes. (b) These contexts are then used to create challenging compositional zero-shot evaluation settings for trajectory prediction, and to enable generalization strategies to enhance OOD robustness.

Our contributions are thus as follows:

1) We introduce a novel, safety-informed scenario factorization approach for autonomous driving, leveraging explicitly disentangled "ego" and "social" axes.
2) We propose new long-tail, zero-shot closed-world and open-world generalization settings upon these axes, where the difficulty-balanced prediction error in the closed-world test setting increases by an average of 5.0% (closed-world) and 14.7% (open-world) over their respective in-distribution performance.
3) We develop generalization techniques that together reduce these OOD performance gaps to 2.8% and 11.5% respectively, eliminating nearly half the gap in closed-world and one quarter in open-world settings, while improving ID performance by 4.0% and 1.2%.

II. Related Work

A. Compositional Zero-Shot Learning

Compositional zero-shot learning (CZSL) is increasingly used in computer vision image-labeling as a framework for evaluating human-like generalization to out-of-distribution examples [9]. In the canonical CZSL setting, models are tasked with labeling novel combinations of semantic factors (i.e., object-attribute pairs) not observed during training [10]. To reduce the generalization gap incurred, various strategies, such as task modular architectures [18], conditional attribute learning [19], and attention propagation [20], have demonstrated promising performance. More recently, approaches leveraging large-scale pre-trained foundation models such as CLIP [21], [22], as well as retrieval-augmented generation (RAG) [23], have achieved substantial improvements therein.

In parallel, these evaluation settings and generalization strategies have also recently been extended from static image-labeling to video *action*-labeling, with verb prompts serving as an analogue to attributes [24], [25]. Nevertheless, our focus on both safety-relevant, semantically meaningful factorizations, as well as behavior *generation* under these conditions rather than just labeling, remains largely unexplored. Furthermore, foundation models are less directly applicable in AD, where the heavy-tailed nature of driving behaviors makes large-scale pre-training alone less effective, especially in distribution shift settings [1], [26].

B. Scenario Characterization and AD Evaluation

For rigorously developing AD systems, scenario characterization and mining approaches are essential, enabling efficient and effective training protocols and more structured evaluation procedures. Identifying scenarios featuring high levels of interaction between traffic participants is particularly common in large dataset curation [15], [16], [27], ensuring that training and evaluation examples are sufficiently challenging. Recent approaches have expanded on this to introduce additional nuanced scenario characteristics, such

as inter-agent causality [28], surprise potential [29], and calibrated regret [30]. UniTraj [31] further proposes a stratified evaluation procedure along Kalman-difficulty bins and agent trajectory shapes. Other approaches have explicitly focused on safety-relevance through automated analyses to capture near-critical scenarios, even while truly safety-critical, long-tail scenarios are absent from recorded datasets [4], [32]. While much of this mining has historically been performed heuristically or with clustering, recent approaches have also explored using large language models and vision language models to identify challenging scenarios and obtain structured scenario descriptors [33]–[35].

Beyond scenario understanding, non-uniform training and testing procedures have also been increasingly utilized. Prior work has explored creating artificial distribution shifts on various bases, including road geometry [36], clustered "domains" stemming from agent trajectory and map statistics [5], [37], and overall safety-relevance [4]. Although some prior work, like [38], has studied compositional OOD generalization along axes like weather and time-of-day, these surface-level axes do not consider the diverse behavior-level attributes that make AD uniquely challenging. Concurrently, many works have explored reducing the observed drops in performance, through techniques like Frenet-based domain normalization [37], collision avoidance losses [4], and closed-loop training with synthetic generated scenarios [39], [40]. However, the entangled nature of such prior domain split techniques makes both performance analysis and broader generalization challenging, with remediation techniques often tailored to specific settings. We instead disentangle scenarios along meaningful, safety-informed axes (i.e., ego and social contexts), and extend modular CZSL techniques to target the gap arising from long-tail compositional challenges.

III. Preliminaries

In this section, we define relevant notation and task definitions for compositional zero-shot evaluation and generalization in trajectory prediction. Concretely, we utilize our pipeline in Section IV-A and Section IV-B to produce discrete "ego" and "social" context labels for agents, though the definitions and notations here are independent of those details.

Scenario and Data Format: Given a particular dataset comprising the set of scenarios $\mathcal{S}$, we denote $\mathbf{s} \in \mathcal{S}$ to refer to a particular scenario therein. Each scenario $\mathbf{s}$ contains agent information for all traffic participants involved, denoted $\mathbf{X}$, as well as map and meta information, denoted $\mathbf{M}$.

We adopt the unified data format and features defined in UniTraj [31]. That is, for a particular agent with id i, $\mathbf{X}_i \in \mathbf{X}$ is a time-indexed feature vector over T timesteps, where $\mathbf{X}_i^{(t)}$ denotes agent i's data at a particular timestep t. This information contains ground-plane kinematics (position, velocity, and acceleration), a valid bit indicating whether the agent is present, as well as one-hot encodings both of the agent's type and the current timestep; i.e., an enriched trajectory representation. Information in $\mathbf{M}$ consists of polyline representations of road and traffic control devices, with locations interpolated at a fixed distance interval and a one-hot encoding of the feature type (e.g., lane, stop-sign, crosswalk, etc.), along with other relevant meta information (e.g., speed limit of lanes, lane adjacency graphs, etc.). Finally, $\mathbf{M}$ also contains a list of "focal" agents that must be motion-predicted, a subset of the total agents in $\mathbf{s}$.

Compositional Zero-Shot Trajectory Prediction: In the standard motion prediction task, time-varying features in $\mathbf{s}$ are split into a *history* and *future*. Given $T = T_{\text{hist}} + T_{\text{fut}}$ representing the total number of timesteps in the scenario, we define $\mathbf{X}_i^{\text{hist}} = \{\mathbf{X}_i^{(t)}\}_{t=1}^{T_{\text{hist}}}$ as the history portion and $\mathbf{X}_i^{\text{fut}} = \{\mathbf{X}_i^{(t)}\}_{t=T_{\text{hist}}+1}^{T}$ as the future portion. The motion prediction task is then to predict the future ground-plane positions in $\mathbf{X}_i^{\text{fut}}$ given only $\mathbf{X}_i^{\text{hist}}$ and $\mathbf{M}$.

For a particular evaluation scheme, the total set of all agents must be split to form a training, validation, and test set. In the prototypical setting, these agents are split uniformly at random according to some desired size of each set, often at the scenario level of granularity (i.e., all focal agents within a given $\mathbf{s} \in \mathcal{S}$ are assigned to the same set). In our CZSL setting, we instead create splits at the agent level, as different focal agents within the same $\mathbf{s}$ may experience very different contexts. We first obtain paired categorical ego and social context labels from all agents across $\mathcal{S}$. We denote these contexts as $\mathcal{C}_{\text{ego}} = \{c_e^k\}_{k=1}^{N_{\text{ego}}}$ and $\mathcal{C}_{\text{social}} = \{c_s^k\}_{k=1}^{N_{\text{social}}}$ for some finite sizes N_{ego} and N_{social}. Thus the compositional class set is defined as the Cartesian product $\mathcal{C} = \mathcal{C}_{\text{ego}} \times \mathcal{C}_{\text{social}}$, following the analogous formulation for image-labeling in [23]. This set is then divided into two subsets, $\mathcal{C}^{\text{seen}}$ for the known compositions, which make up the training and validation set, and $\mathcal{C}^{\text{unseen}}$ for the unknown compositions that make up the test set. Importantly, there is no overlap between $\mathcal{C}^{\text{seen}}$ and $\mathcal{C}^{\text{unseen}}$, meaning that agent examples in the test set are out-of-distribution with respect to the training set.

In the *closed-world* test setting, $\mathcal{C}^{\text{unseen}}$ consists solely of novel *combinations* of known ego and social contexts; that is, for each paired $(c_e, c_s) \in \mathcal{C}^{\text{unseen}}$, both c_e and c_s are present in $\mathcal{C}^{\text{seen}}$, but never jointly on the same example. In the harder *open-world* test setting, however, we require that for each joint $(c_e, c_s) \in \mathcal{C}^{\text{unseen}}$, at least one of c_e or c_s is completely absent from $\mathcal{C}^{\text{seen}}$. Thus, the task of compositional zero-shot trajectory prediction requires developing a model that performs well on both $\mathcal{C}^{\text{seen}}$ and $\mathcal{C}^{\text{unseen}}$, despite training only on $\mathcal{C}^{\text{seen}}$. Unlike conventional CZSL, which requires *predicting labels* for unseen compositions, our setting requires *generating future behavior* for agents in unseen compositional contexts $(c_e, c_s) \in \mathcal{C}^{\text{unseen}}$.

IV. Approach

To both construct long-tail, safety-relevant compositional settings for trajectory prediction and enhance generalization performance therein, we first extract relevant ego and social features (Section IV-A). We then discretize these features into context labels (Section IV-B) and use them to create challenging train/test splits (Section IV-C). Finally, we extend baseline trajectory prediction models with new generalization

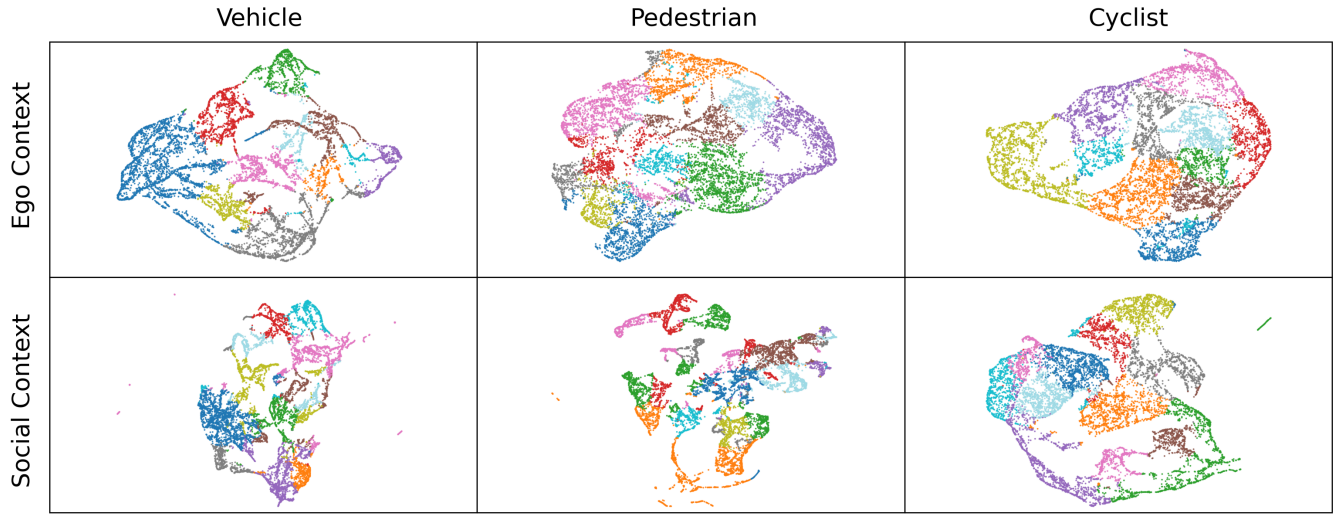


Fig. 2: UMAP [41] visualizations for ego and social contexts across agent types. Colors correspond to discretized context labels from clustering described in Section IV-B, independently per diagram.

modules to improve out-of-distribution (OOD) performance (Section IV-D).

A. Safety-Relevant Feature Extraction

We begin by extending prior work on scenario characterization, namely, SafeShift [4] and IMGTP [5], deriving broader sets of both safety-relevant and general attributes and systematically processing them to support downstream discretization. Given a focal agent i in scenario $\mathbf{s}$, we first linearly interpolate $\mathbf{X}_i^{\text{hist}}$ and each corresponding $\mathbf{X}_{k\neq i}^{\text{hist}}$ over agent i's valid timestep bounds. We utilize only information from the *history* portion of $\mathbf{s}$ to avoid any leakage from ground-truth future trajectories. We then compute and process the following features:

Ego features: These focus on kinematic details, lane information, and traffic control device proximity. We first extract agent i's position, velocity, acceleration, and curvature, relative to its final pose (i.e., its position and heading at T_{hist}). Then, if a valid lane assignment exists based on heading similarity and lateral distance thresholds, we include speed-limit compliance, lane type, and lane-relative, Frenet coordinates of agent i's trajectory. Finally, we obtain distances and relative headings to stop signs, crosswalks, traffic lights, and speed bumps, if such infrastructure is present.

Social features: These instead focus on agent i's interactions with other relevant agents. We start by computing global scalar attributes, like scenario density. Then, for each non-stationary external agent within a given distance threshold to the focal agent, we compute a *geometry* type of the interaction at T_{hist}. We first assign interactions into collinear, parallel, opposite, or crossing types, using longitudinal and lateral distances, as well as relative heading differences. We further break down collinear geometries to leading, trailing, or head-on variants, and all other geometries to left or right variants relative to agent i.

Next, we compute time-varying kinematic differences to the focal agent, both relative to agent i's final pose as well as to each of i's poses in $\mathbf{X}_i^{\text{hist}}$. We explicitly capture collision-relevance of the interaction via closing speed and linearly projected conflict points at a fixed future horizon (i.e., projected locations of closest separation). For such conflict points, we compute the distance and pose of the point itself, as well as the difference in time-to-conflict-point ($\Delta TTCP$, established in [42]). Finally, we include a categorical feature capturing the object type (i.e., vehicle, pedestrian, or cyclist) of the external agent.

B. Feature Processing and Context Discretization

A key challenge in discretizing the above ego and social features into (c_e, c_s) context labels is that scenarios contain variable numbers and types of agents, traffic control devices, and other scenario elements, even including cases where none are present. We, therefore, process the extracted features into consistent, fixed-length representations, better supporting autoencoding and clustering.

We first organize input features into co-occurring groups. For ego features, these groups include the focal agent's kinematics, lane information, the closest instance of each traffic control device type, and the closest instance of each type located ahead of agent i. For social features, these groups include global features and the closest external agent of each geometry type. Within each group, we one-hot encode categorical values, summarize time-varying numerical values with scalar statistics (i.e., mean, standard deviation, minimum and maximum, and average slope), and directly include static numerical values. We then concatenate the features from all groups into a consistently ordered vector for each axis; if any group is absent (e.g., there is no forward stop-sign relative to agent i, or no agent trailing agent i collinearly, etc.), we zero-fill the required features. Finally, we append a valid bit to each group indicating whether or not it was present, producing the final ego and social vectors v_e and v_s.

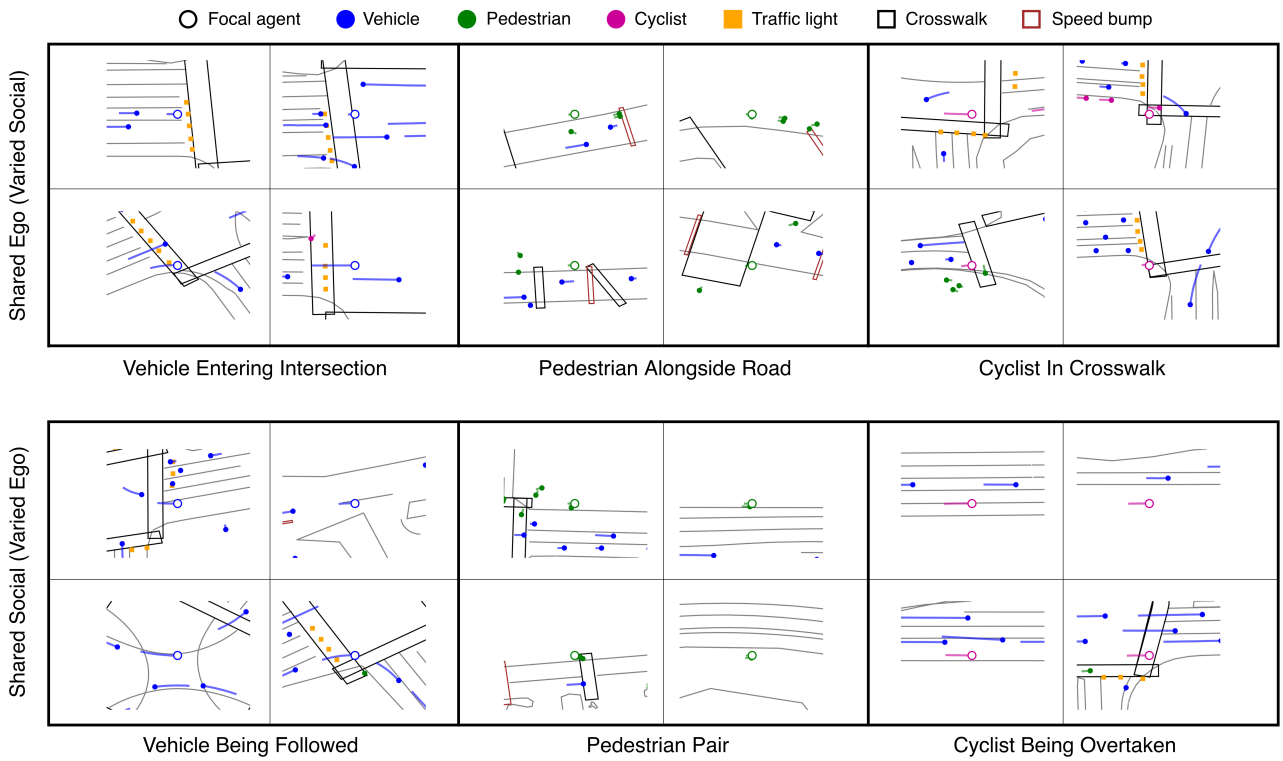


Fig. 3: Cluster examples, by context and agent behavior types. In each subgrid of 4 examples, one context (e.g., ego) is fixed while the paired context (e.g., social) varies, with a brief caption describing the shared behavior. Agent markers show positions at T_{hist}; the focal agent is shown as a large, open circle, while background agents are smaller and filled.

Next, we train autoencoders for ego and social vectors separately. We further split by object type and train independent models for each type, allowing distinct latent spaces to be learned for e.g., pedestrian focal agents versus vehicle focal agents. Each autoencoder consists of a simple encoder and decoder multi-layer perceptron (MLP), with layer normalization and dropout on hidden layers; the encoder maps down to a low-dimensional latent space and the decoder maps back to the original feature space. That is, we compute $z = \text{Enc}(v)$ and $\tilde{v} = \text{Dec}(z)$. We train the models primarily with a mean-square error (MSE) reconstruction loss between v and $\tilde{v}$, along with a deep embedding clustering (DEC) [43] loss for regularization on the latent z values.

We then obtain discrete ego and social contexts by performing clustering within the latent spaces captured by these autoencoders, using k-means with $k = 11$. We use the Waymo Open Motion Dataset (WOMD) [15] as a representative source of AD scenarios, sampling approximately 20% of the total data. To quantitatively assess cluster and latent space coherence, we compute silhouette scores on held-out sets [44], observing values ranging from 0.31 to 0.50, which indicates a reasonably well-structured space. We also visualize UMAP [41] projections of the resulting spaces in Figure 2, showing clear separation and evidence of potential sub-clusters. We further present cluster examples for each latent space in Figure 3, demonstrating intra-cluster consistency alongside paired-context diversity. Ultimately, these intermediate results confirm the validity of our ego and social context derivations, supporting their use as c_e and c_s in downstream compositional tasks.

C. Closed-World and Open-World Settings

To ensure that held out context combinations indeed contain sufficient long-tail behaviors, we follow UniTraj [31] in considering Kalman difficulty as a reasonably well-calibrated approximation for the overall trajectory prediction challenge. That is, we compute the final displacement error a linear Kalman filter incurs when forecasting the focal agent i, averaged at fixed time horizons (i.e., 2, 4, and 6 seconds in the future). We then average this difficulty value for each clustered context obtained in Section IV-B.

Given these context difficulty values, we construct our *open-world* setting by greedily holding out all agents in the highest average-difficulty ego context (regardless of paired social context), as well as those in the highest average-difficulty social context (regardless of paired ego context), repeating this process until a desired test-set size is achieved. For our *closed-world* setting, we follow the same process, but for a given ego or social context that was held out, we add back in examples from *half* of its co-occurring paired contexts. That is, if e.g., all of c_e was originally held out, we add back $(c_e, c_s^{\{1,2,\ldots,N_{\text{social}}/2\}})$.

In both settings, the remaining examples are then split uniformly into training and validation sets of desired sizes. Again using WOMD [15] as a representative dataset, we observe that the average Kalman difficulties in the created

test sets are both approximately 20% higher than their corresponding validation set difficulties. This similar magnitude of increase thus allows us to examine the impact of the *compositional* differences between the two long-tail settings, when used in downstream motion prediction experiments.

D. Generalization Strategies

We propose two techniques for improving performance in the above challenging settings: extending task-modular networks (TMNs) [18] for behavior generation, and refining intermediate representations through a difficulty-prediction auxiliary objective. Given a prototypical encoder-decoder trajectory prediction model, we operate directly in its bottleneck layer for both of these ideas. That is, given an intermediate representation of a focal agent as h, we transform it into an enriched h' before passing it to the model's decoder.

As in the original TMN, the intuition is that a gating network produces weightings to leverage learned modules in novel ways, as appropriate for novel context combinations. However, whereas the original TMN approach ultimately yields a compatibility score between various candidate context labels and the input representation, we refine the original h representation conditioned on fine-grained, sample-level latents.

Our TMN-inspired architecture thus consists of two main components: a sequence of N layers of M MLP "modules", and a latent-conditioned gating MLP $\mathcal{G}$. The first module layer processes M copies of the initial h vector, while modules in subsequent layers process weighted sums of the previous layer's outputs; these weights are obtained from the gating network's output. A linear projection head then maps the final module layer back to the original input space, producing h'.

To help ensure that such weightings from $\mathcal{G}$ are coherent, even in open-world settings, we condition $\mathcal{G}$ on the structured latent spaces obtained via the autoencoding process described in Section IV-B. Note that to avoid information leakage from the test split, we retrain these autoencoders on the closed-world and open-world train sets described in Section IV-C, producing z'_e and z'_s vectors that can safely be used as input to $\mathcal{G}$. We train this component jointly with the baseline prediction model's objectives.

We additionally propose to further refine this h' by enhancing it with difficulty-awareness. We add an auxiliary head, formulated as a simple linear layer on top of h', to map to three scalar values, representing the anticipated Kalman error at 2, 4, and 6 seconds (trained with an MSE regression loss). In this way, we help promote implicit reasoning over difficulty-conditioned decision making (e.g., cautious versus aggressive behavior, etc.), which is especially helpful in our long-tail, safety-informed settings.

V. Experiments

Dataset and training details: As in our intermediate results in Section IV, we use the Waymo Open Motion Dataset (WOMD) [15] as our high quality source of scenarios, processed into a standardized ScenarioNet [45] format. We sample approximately 500k agent trajectories from these scenarios to perform ego and social context derivation, evaluation setting creation, and trajectory prediction experiments therein. As in the standard WOMD setting, we set T_{hist} to 11 and T_{fut} to 80 timesteps, at 10 Hz.

In the characterization process, we use a heading alignment threshold of 30 degrees and a lateral threshold of 6.5m when considering lane assignments. We additionally set an overall distance threshold of 50m for filtering out irrelevant agents, and use relative heading thresholds of 30 degrees and a collinear lateral threshold of 3.25m when determining geometry types. We also use a 10 second horizon when considering linearly projected conflict points.

When training autoencoders for context discretization, we use a latent dimension of $z = 16$, compressing and reconstructing v_e and v_s with input dimensions of 346 and 1443, resulting in models with approximately 500k and 800k parameters, respectively. We then use $k = 11$ clusters both for the DEC loss module and the k-means computation.

For conducting trajectory prediction experiments, we use the well-established Motion Transformer (MTR) [17] model, scaled to approximately 2.5M total parameters. In our generalization strategies, we utilize $N = 3$ intermediate layers and $M = 12$ modules per layer, where the gating network conditions on both z'_e and z'_s, with dimensions of 16 each. Overall, we add fewer than 100k new parameters to the base MTR model, across both ideas in Section IV-D. For finer-grained analyses, we train MTR augmented with the TMN-inspired component and auxiliary head jointly ("MTR + Both"), as well as with each component added separately as ablations ("MTR + TMN-Inspired" and "MTR + Auxiliary"). We train all models for 30 epochs, with an initial learning rate of $1e-3$, a fixed decay schedule, and a batch size of 192.

Metrics and evaluation: We use standard motion forecasting metrics, focusing on displacement differences between ground truth positions in X^{fut} and predicted trajectories from the above models. Since future motion is multi-modal, typical models produce multiple distinct future trajectories along with confidence or probability values that a particular mode is best; we set the number of modes to six, as in WOMD [15] and Argoverse [16]. Then, we utilize the minimum Average and Final Displacement Errors (ADE and FDE) to capture L2 distances over all valid T_{fut} timesteps, as well as just at the final observed T, for the mode which best matches the ground truth. We additionally report Brier-FDE values, which penalizes FDE by $(1-p)^2$, where p is the confidence in the best mode, as popularized in the Argoverse challenge.

Although the closed-world and open-world test sets in Section IV-C contain more examples of difficult, long-tail behavior than their corresponding validation sets, the distribution still remains imbalanced with an over-representation of "easy" cases. To better isolate the impacts of the zero-shot compositional challenges, as well as the efficacy of our generalization strategies, we thus report all metrics averaged over the Kalman difficulty-class stratification established in

TABLE I: Generalization results for the compositional settings; `Seen` results are in-distribution while `Unseen` results are out-of-distribution to the train set. Numbers in parentheses indicate relative change from the MTR baseline value in `Seen` (i.e., the setting's generalization gap on that metric). **Lower** numbers are better for all metric values and relative changes.

Setting	Method	Seen ADE	Seen FDE	Seen Brier-FDE	Unseen ADE	Unseen FDE	Unseen Brier-FDE
Closed-World	MTR	2.11 (−)	4.63 (−)	5.12 (−)	2.11 (+0.2%)	4.99 (+7.8%)	5.47 (+6.9%)
	MTR + TMN-Inspired	2.11 (−0.1%)	4.63 (−0.1%)	5.10 (−0.4%)	2.10 (−0.6%)	4.94 (+6.8%)	5.42 (+5.9%)
	MTR + Auxiliary	2.06 (−2.6%)	**4.37** (**−5.6%**)	**4.85** (**−5.2%**)	**2.08** (**−1.5%**)	4.85 (+4.7%)	5.33 (+4.1%)
	MTR + Both	**2.05** (**−2.8%**)	4.41 (−4.7%)	4.89 (−4.5%)	2.11 (+0.1%)	**4.84** (**+4.5%**)	**5.32** (**+3.9%**)
Open-World	MTR	1.97 (−)	4.33 (−)	4.82 (−)	2.12 (+7.8%)	5.15 (+19.0%)	5.65 (+17.4%)
	MTR + TMN-Inspired	2.00 (+1.7%)	4.41 (+1.8%)	4.89 (+1.6%)	2.12 (+7.5%)	5.13 (+18.5%)	5.62 (+16.8%)
	MTR + Auxiliary	2.07 (+5.1%)	4.52 (+4.5%)	5.01 (+4.1%)	2.14 (+8.6%)	5.17 (+19.5%)	5.67 (+17.7%)
	MTR + Both	**1.94** (**−1.5%**)	**4.28** (**−1.0%**)	**4.77** (**−1.0%**)	**2.06** (**+4.8%**)	**5.01** (**+15.6%**)	**5.49** (**+14.1%**)

UniTraj [31]. Finally, for all experiments, we select the best performing validation Brier-FDE checkpoint and use it to conduct inference on the held-out test sets, obtaining performance measures on both in-distribution compositions in $\mathcal{C}^{\text{seen}}$ and out-of-distribution compositions in $\mathcal{C}^{\text{unseen}}$.

VI. Results

We present our main experiment results in Table I. To understand the severity of our long-tail compositional zero-shot settings, we first report baseline metric values with MTR alone. As in SafeShift [4], we report each metric along with its relative change from this MTR baseline `Seen` value; error values increase by an average of $\mathbf{5.0}\%$ and $\mathbf{14.7}\%$ in closed-world and open-world settings, respectively. Note that in both settings, the drop in ADE performance is less than FDE, since shorter T_{fut} horizons tend to be easier to predict.

Overall, these results confirm that both our closed-world and open-world zero-shot settings indeed induce significant drops in OOD performance, even for a state-of-the-art model like MTR. Importantly, despite the fact that the relative Kalman difficulty between `Seen` and `Unseen` sets is similar in both settings, the open-world drop in performance was *far* larger than the closed-world drop, clearly demonstrating the impact of compositional setting design, as discussed in Section IV-C.

We then assess our generalization strategies proposed in Section IV-D. While the TMN-inspired and auxiliary loss components alone improve certain metrics in our settings, using both components achieves the strongest and most consistent performance, highlighting the efficacy of their joint refinement of the hidden representation space. In particular, our full approach reduces the `Unseen` performance gap to an average of $\mathbf{2.8}\%$ and $\mathbf{11.5}\%$ in closed-world and open-world, respectively. Furthermore, performance in the `Seen` setting improved by an average of 4.0% in closed-world and 1.2% in open-world. Hence, our proposed strategy of combining modular reasoning using ego and social representations, along with implicitly considering the difficulty of a given example, improves performance broadly, but especially in the OOD `Unseen` sets.

VII. Conclusion

Robust development and evaluation of trajectory prediction models remain an essential challenge in autonomous driving, especially in the presence of long-tail, safety-critical scenarios encountered in deployment, which are rare or missing altogether in large-scale datasets. We therefore proposed novel evaluation settings on existing datasets, where test sets are constructed to be OOD with respect to training data on the bases of safety-informed context compositions—to our knowledge, the **first** extension of the well-established image-labeling CZSL setting to motion prediction. To create these settings, we developed a factorization framework that disentangles traffic participants along ego and social axes, clustering them into paired context labels. We then used these contexts to build long-tail, zero-shot closed-world and open-world settings. On a state-of-the-art baseline model, we observed performance drops from ID data of 5.0% and 14.7%, respectively; by extending task-modular gating networks and incorporating an auxiliary, difficulty-prediction loss, we reduced these OOD gaps to 2.8% and 11.5%, while also improving ID performance by 4.0% and 1.2%, respectively.

Despite the efficacy of our evaluation settings and generalization strategies, further improvements are still possible. In particular, extending other non CLIP-based approaches from image-labeling CZSL, such as learned attention propagation [20], could boost both ID and OOD performance. Additionally, factorizing along more than two safety-relevant, semantic axes could also increase interpretability, as well as the effectiveness of generalization approaches. We encourage future work to explore these directions.

References

[1] H. X. Liu and S. Feng, "Curse of rarity for autonomous vehicles," *nature communications*, vol. 15, no. 1, p. 4808, 2024.

[2] Z. Wang, J. Guo, H. Zhang, R. Wan, J. Zhang, and J. Pu, "Bridging the gap: Improving domain generalization in trajectory prediction," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 1, pp. 1780–1791, 2023.

[3] W. Ding, C. Xu, M. Arief, H. Lin, B. Li, and D. Zhao, "A survey on safety-critical driving scenario generation—a methodological perspective," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 7, pp. 6971–6988, 2023.

[4] B. Stoler, I. Navarro, M. Jana, S. Hwang, J. Francis, and J. Oh, "Safeshift: Safety-informed distribution shifts for robust trajectory prediction in autonomous driving," in *2024 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2024, pp. 1179–1186.

[5] L. Ye, Z. Zhou, J. Wang, Y.-H. Li, N.-Y. Jan, Y.-R. Lin, and Y.-C. Lin, "Imgtp: A unified framework for improving and measuring the generalizability of trajectory prediction models," *IEEE Transactions on Intelligent Vehicles*, 2024.

[6] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 8, pp. 1798–1828, 2013.

[7] X. Wang, H. Chen, Z. Wu, W. Zhu, *et al.*, "Disentangled representation learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[8] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, "Building machines that learn and think like people," *Behavioral and brain sciences*, vol. 40, p. e253, 2017.

[9] F. Pourpanah, M. Abdar, Y. Luo, X. Zhou, R. Wang, C. P. Lim, X.-Z. Wang, and Q. J. Wu, "A review of generalized zero-shot learning methods," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 4, pp. 4051–4070, 2022.

[10] M. Mancini, M. F. Naeem, Y. Xian, and Z. Akata, "Open world compositional zero-shot learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 5222–5230.

[11] Y. Atzmon, F. Kreuk, U. Shalit, and G. Chechik, "A causal view of compositional zero-shot recognition," *Advances in Neural Information Processing Systems*, vol. 33, pp. 1462–1473, 2020.

[12] N. Medeiros-Ward, J. M. Cooper, and D. L. Strayer, "Hierarchical control and driving." *Journal of experimental psychology: General*, vol. 143, no. 3, p. 953, 2014.

[13] N. A. Stanton and P. M. Salmon, "Human error taxonomies applied to driving: A generic driver error taxonomy and its implications for intelligent transport systems," *Safety Science*, vol. 47, no. 2, pp. 227–237, 2009.

[14] S. G. Charlton, N. J. Starkey, J. A. Perrone, and R. B. Isler, "What's the risk? a comparison of actual and perceived driving risk," *Transportation research part F: traffic psychology and Behaviour*, vol. 25, pp. 50–64, 2014.

[15] S. Ettinger, S. Cheng, B. Caine, C. Liu, H. Zhao, S. Pradhan, Y. Chai, B. Sapp, C. R. Qi, Y. Zhou, *et al.*, "Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 9710–9719.

[16] 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: Next generation datasets for self-driving perception and forecasting," *arXiv preprint arXiv:2301.00493*, 2023.

[17] S. Shi, L. Jiang, D. Dai, and B. Schiele, "Motion transformer with global intention localization and local movement refinement," *Advances in Neural Information Processing Systems*, vol. 35, pp. 6531–6543, 2022.

[18] S. Purushwalkam, M. Nickel, A. Gupta, and M. Ranzato, "Task-driven modular networks for zero-shot compositional learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 3593–3602.

[19] Q. Wang, L. Liu, C. Jing, H. Chen, G. Liang, P. Wang, and C. Shen, "Learning conditional attributes for compositional zero-shot learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 11 197–11 206.

[20] M. G. Z. A. Khan, M. F. Naeem, L. Van Gool, A. Pagani, D. Stricker, and M. Z. Afzal, "Learning attention propagation for compositional zero-shot learning," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, 2023, pp. 3828–3837.

[21] N. V. Nayak, P. Yu, and S. H. Bach, "Learning to compose soft prompts for compositional zero-shot learning," in *International Conference on Learning Representations*, 2023.

[22] W. Bao, L. Chen, H. Huang, and Y. Kong, "Prompting language-informed distribution for compositional zero-shot learning," in *European Conference on Computer Vision*. Springer, 2024, pp. 107–123.

[23] C. Jing, Y. Li, H. Chen, and C. Shen, "Retrieval-augmented primitive representations for compositional zero-shot learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 38, no. 3, 2024, pp. 2652–2660.

[24] R. Li, Z. Feng, T. Xu, L. Li, X.-J. Wu, M. Awais, S. Atito, and J. Kittler, "C2c: Component-to-composition learning for zero-shot compositional action recognition," in *European Conference on Computer Vision*. Springer, 2024, pp. 369–388.

[25] G. Ye, L. Li, K. Li, J. Xiao, *et al.*, "Zero-shot compositional action recognition with neural logic constraints," *arXiv preprint arXiv:2508.02320*, 2025.

[26] W. Dai, S. Wu, W. Wu, Z. Wang, S. Lyu, H. Liao, L. Yu, W. Ding, R. Guan, and Y. Yue, "Large foundation models for trajectory prediction in autonomous driving: A comprehensive survey," *arXiv preprint arXiv:2509.10570*, 2025.

[27] T. Moers, L. Vater, R. Krajewski, J. Bock, A. Zlocki, and L. Eckstein, "The exid dataset: A real-world trajectory dataset of highly interactive highway scenarios in germany," in *2022 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2022, pp. 958–964.

[28] R. Roelofs, L. Sun, B. Caine, K. S. Refaat, B. Sapp, S. Ettinger, and W. Chai, "Causalagents: A robustness benchmark for motion forecasting using causal relationships," *arXiv preprint arXiv:2207.03586*, 2022.

[29] W. Ding, S. Veer, K. Leung, Y. Cao, and M. Pavone, "Surprise potential as a measure of interactivity in driving scenarios," *arXiv preprint arXiv:2502.05677*, 2025.

[30] K. Nakamura, R. Tian, and A. Bajcsy, "A general calibrated regret metric for detecting and mitigating human-robot interaction failures." *CoRR*, 2024.

[31] L. Feng, M. Bahari, K. M. B. Amor, É. Zablocki, M. Cord, and A. Alahi, "Unitraj: A unified framework for scalable vehicle trajectory prediction," in *European Conference on Computer Vision*. Springer, 2024, pp. 106–123.

[32] C. Glasmacher, R. Krajewski, and L. Eckstein, "An automated analysis framework for trajectory datasets," *arXiv preprint arXiv:2202.07438*, 2022.

[33] Y. Yang, Q. Zhang, K. Ikemura, N. Batool, and J. Folkesson, "Hard cases detection in motion prediction by vision-language foundation models," in *2024 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2024, pp. 2405–2412.

[34] S. Tan, B. Ivanovic, X. Weng, M. Pavone, and P. Kraehenbuehl, "Language conditioned traffic generation," in *7th Annual Conference on Robot Learning*, 2023. [Online]. Available: https://openreview.net/forum?id=PK2debCKaG

[35] X. Zheng, L. Wu, Z. Yan, Y. Tang, H. Zhao, C. Zhong, B. Chen, and J. Gong, "Large language models powered context-aware motion prediction in autonomous driving," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 980–985.

[36] A. Filos, P. Tigkas, R. McAllister, N. Rhinehart, S. Levine, and Y. Gal, "Can autonomous vehicles identify, recover from, and adapt to distribution shifts?" in *International Conference on Machine Learning*. PMLR, 2020, pp. 3145–3153.

[37] L. Ye, Z. Zhou, and J. Wang, "Improving the generalizability of trajectory prediction models with frenet-based domain normalization," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 11 562–11 568.

[38] T.-H. Wang, A. Maalouf, W. Xiao, Y. Ban, A. Amini, G. Rosman, S. Karaman, and D. Rus, "Drive anywhere: Generalizable end-to-end autonomous driving with multi-modal foundation models," in *2024 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2024, pp. 6687–6694.

[39] L. Zhang, Z. Peng, Q. Li, and B. Zhou, "Cat: Closed-loop adversarial training for safe end-to-end driving," in *Conference on Robot Learning*. PMLR, 2023, pp. 2357–2372.

[40] B. Stoler, I. Navarro, J. Francis, and J. Oh, "Seal: Towards safe autonomous driving via skill-enabled adversary learning for closed-loop scenario generation," *arXiv preprint arXiv:2409.10320*, 2024, under review for ICRA 2025.

[41] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[42] W. Zhan, L. Sun, D. Wang, Y. Jin, and M. Tomizuka, "Constructing a highly interactive vehicle motion dataset," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2019, pp. 6415–6420.

[43] J. Xie, R. Girshick, and A. Farhadi, "Unsupervised deep embedding for clustering analysis," in *International conference on machine learning*. PMLR, 2016, pp. 478–487.

[44] K. R. Shahapure and C. Nicholas, "Cluster quality analysis using silhouette score," in *2020 IEEE 7th international conference on data science and advanced analytics (DSAA)*. IEEE, 2020, pp. 747–748.

[45] Q. Li, Z. M. Peng, L. Feng, Z. Liu, C. Duan, W. Mo, and B. Zhou, "Scenarionet: Open-source platform for large-scale traffic scenario simulation and modeling," *Advances in neural information processing systems*, vol. 36, pp. 3894–3920, 2023.