

Trajectory Prediction for Autonomous Driving in Road Construction Zones Using Human-in-the-Loop Driving Data

Shuo Zhang*
sz4580@nyu.edu
New York University
New York City, NY, USA

Semiha Ergan
semiha@nyu.edu
New York University
New York, USA

Aman Sunesh*
s18181@nyu.edu
New York University
New York City, NY, USA

Kaan Ozbay
kaan.ozbay@nyu.edu
New York University
New York City, NY, USA

Abstract

Ego-trajectory prediction in urban roadway work zones remains an underexplored topic in autonomous-driving research, with limited dedicated data and systematic evaluation. Roadway work zones involve temporary changes in the lane layouts and connectivity of the roads, complicating interactions with surrounding vehicles and workers. These changes make future vehicle trajectories difficult to predict. To address these challenges, we collect WZDRIVE-DATA, a multimodal dataset comprising 1,183 work-zone driving segments from 33 drivers, using a human-in-the-loop driving simulator with prerecorded worker animations and VR-controlled workers. Driver control and gaze signals are recorded from human participants, while vehicle trajectories and scene information are obtained from the simulator. Our proposed multimodal ego-trajectory predictor, WZDRIVE, achieves a mean minADE₆ of 1.09 m and minFDE₆ of 2.12 m. Compared with the best-performing state-of-the-art baseline for each metric in our benchmark, WZDrive reduces minADE₆ by 7.18% and minFDE₆ by 13.18%. This research addresses a gap in autonomous-driving research on urban work zones by combining human-in-the-loop data collection with multimodal learning from human driving signals and in the context of work-zones. A project website with code and example data is available at <https://shuo9898-zs.github.io/wzdrive/>.

Keywords

Ego-trajectory prediction, Autonomous driving, Roadway work zones, Multimodal learning, Human-in-the-loop simulation

1 Introduction

Roadway work zones are essential for maintaining and upgrading transportation and utility infrastructure, but they temporarily change the conditions under which vehicles operate. Maintenance activities can narrow or redirect travel lanes and require drivers to respond to surrounding vehicles, temporary traffic-control devices, and roadway workers in constrained spaces. These temporary changes can create safety-critical conditions for drivers and workers. In the United States, an estimated 101,000 work-zone crashes occurred in 2023, resulting in roughly 39,000 injuries and about 900 fatalities [1, 2]. As aging infrastructure and continuing utility work sustain the need for maintenance and rehabilitation [3], roadway

work zones will remain a recurring part of road travel and an important setting for studying how vehicles adapt to temporary road configurations.

Ego-trajectory prediction estimates a vehicle’s near-future motion to support autonomous driving (AD) and advanced driver-assistance systems (ADAS). In work zones, these predictions can help AD systems assess where the vehicle is likely to travel and plan safer paths around workers and nearby traffic. For ADAS, they can support hazard warnings and corrective control when a driver’s actions may lead to unsafe motion. However, temporary lane changes and interactions with vehicles and workers make prediction challenging. Limited road space in urban areas can restrict standard traffic-control layouts, while short-term work zones may provide less warning and protection, leaving drivers less time to respond.

Dedicated work-zone datasets have advanced scene understanding and navigation. EPOSH focuses on perspective-view and bird’s-eye-view segmentation [4], while ROADWork provides rich visual annotations and reconstructed driving paths [5]. However, ROADWork’s pathway-prediction task uses spatially sampled waypoints derived from video-based reconstruction rather than directly recorded vehicle states at fixed time intervals. General driving datasets, including nuScenes, Argoverse, and the Waymo Open Motion Dataset, provide temporal trajectories and established forecasting benchmarks [6–8]. Yet their collection and evaluation protocols are not specifically organized around safety-critical work-zone conditions. This leaves a need for dedicated trajectory data and systematic evaluation of vehicle motion under these conditions.

Human signals offer a potential means of addressing prediction challenges that trajectory history alone cannot resolve. Prior work has developed immersive platforms for collecting human driving signals [9] and explored EEG-based feedback and attention-based guidance for ADAS and autonomous driving [10–12]. These studies motivate investigating human signals as complementary information for predicting driving responses in complex work-zone conditions. We therefore propose a multimodal approach that combines historical driver-control and attention signals with vehicle trajectories, surrounding agents, lane geometry, and temporary work-zone context. To address ego-trajectory prediction in urban roadway work zones, we recruited 33 participants to collect driving data

*Shuo Zhang and Aman Sunesh contributed equally to this work.

across six scenarios using a CARLA–SUMO-based driving simulator with VR-controlled workers and prerecorded worker animations [13, 14].

Our contributions includes:

- **Dataset:** We collect WZDRIVE-DATA, comprising 1,183 work-zone driving segments with synchronized vehicle trajectories, driver signals, and scene context across six scenario conditions.
- **Benchmark:** We benchmark rule-based, sequence, and state-of-the-art models for ego-trajectory prediction in roadway work zones.
- **Model:** We propose WZDRIVE, which combines multimodal encoding and stage-wise refinement to improve ego-trajectory prediction over the evaluated baselines.

2 Related Work

2.1 Trajectory Prediction

Trajectory prediction supports autonomous driving and navigation among human road users by estimating future motion from observed histories and scene context. Trajectron++ models multimodal agent behavior [15], while Social-STGCNN captures pedestrian interactions through spatiotemporal graphs [16]. Map-aware approaches incorporate lane structure and agent interactions through lane graphs, graph traversals, and hierarchical representations [17–19].

Attention-based models jointly represent agents, time, and heterogeneous scene inputs [20, 21]. MTR, MTR++, and QCNet use query-based prediction and refinement [22–24], while Forecast-MAE pretrains representations through masked trajectory and lane reconstruction [25]. LAFormer combines lane-aware constraints with two-stage refinement [26], and DeMo separates directional intentions from dynamic states [27]. For ego-trajectory prediction, RouteFormer incorporates trajectory history, egocentric imagery, and driver field of view [28]. These methods primarily address general driving environments rather than the joint use of temporary work-zone geometry and human-driver control signals.

2.2 Driving and Roadway-Work-Zone Datasets

General driving datasets, including nuScenes, Argoverse, and the Waymo Open Motion Dataset, provide trajectories and map context for motion forecasting [6–8]. Driver-centered datasets additionally capture human behavior and attention. DR(eye)VE includes driving video, vehicle states, and gaze [29], while GEM, introduced with RouteFormer, synchronizes egocentric video, corrected ego locations, and gaze [28]. Zhuang et al. collect EEG, gaze, and driver-control data in VR driving simulations to support human-informed autonomous vehicle control [10]. These resources support complementary aspects of driving research but do not specifically target human driving responses in roadway work zones.

Work-zone research spans perception, scene understanding, and navigation. Early work identified work-zone boundaries and traffic changes from signs [30]. EPOSH provides perspective-view and bird’s-eye-view semantic annotations [4]. ROADWork adds object and sign annotations, scene descriptions, and reconstructed driving paths [5]. More recently, Dash2Sim reconstructs metric,

geo-referenced 4D driving logs from dashcam videos and introduces ROADWork4D-CL for closed-loop planning evaluation [31]. Our dataset complements these resources by pairing controlled human-driving trajectories with synchronized driver signals and explicit work-zone geometry for ego-trajectory prediction.

2.3 Human-in-the-Loop Simulation for Road-User Interaction

Immersive simulation enables controlled collection of interactions that would be difficult or unsafe to reproduce in real traffic. Weng et al. evaluated VR-based collection of pedestrian trajectories and body poses in safety-critical interactions [32]. JaywalkerVR and its CARLA-VR dataset further support controlled studies of pedestrian behavior [33]. Related VR frameworks investigate human-driver–pedestrian interactions, trajectory and crossing prediction, and coupled intentions and actions [34, 35]. Dang et al. additionally model bounded rationality in game-theoretic vehicle–pedestrian decision making [36].

Driver-centered platforms demonstrate the feasibility of immersive simulation for multimodal human-driving research. DReyeVR integrates CARLA with VR driving, eye tracking, driver controls, and experiment replay [9]. DualTake combines immersive simulation with gaze, physiological, and vehicle signals to predict takeover behavior across cars and micromobility [37]. For roadway work zones, Zhang et al. combine VR–CARLA simulation with eye tracking to study driver responses [38], while VR-WISE integrates CARLA–SUMO co-simulation with gaze and physiological sensing to investigate driver awareness [39]. These studies provide a methodological basis for collecting human responses under controlled driving conditions.

Work zones are not inherently safety-critical, but particular layouts, environmental conditions, and worker interactions can increase conflict potential. This motivates a trajectory-centered benchmark that combines human-driving signals and the context of both work-zone under a matched control condition and crash-informed, safety-critical configurations.

3 Dataset

3.1 Human-in-the-Loop Data Collection

Data collection platform. We collected WZDRIVE-DATA using a human-in-the-loop platform integrating CARLA–SUMO co-simulation [13, 14], physical driving controls, gaze tracking, scene sensing, and a VR interface for roadway-worker interaction (Fig. 1). The platform records synchronized human-driving and scene data for ego-trajectory prediction in roadway work zones. The collection platform is described on the DuoRoad project website.¹

Work-zone and scenario design. The study included three work-zone configurations. WZ1 and WZ3 used prerecorded worker animations within the work-zone boundary. In WZ2, a human VR operator placed traffic cones to form a temporary boundary while interacting with the driver in real time. Three students with civil-engineering backgrounds served as VR-worker operators, enabling dual-human-in-the-loop interaction. Each participant encountered

¹<https://shuo9898-zs.github.io/duoroad/>

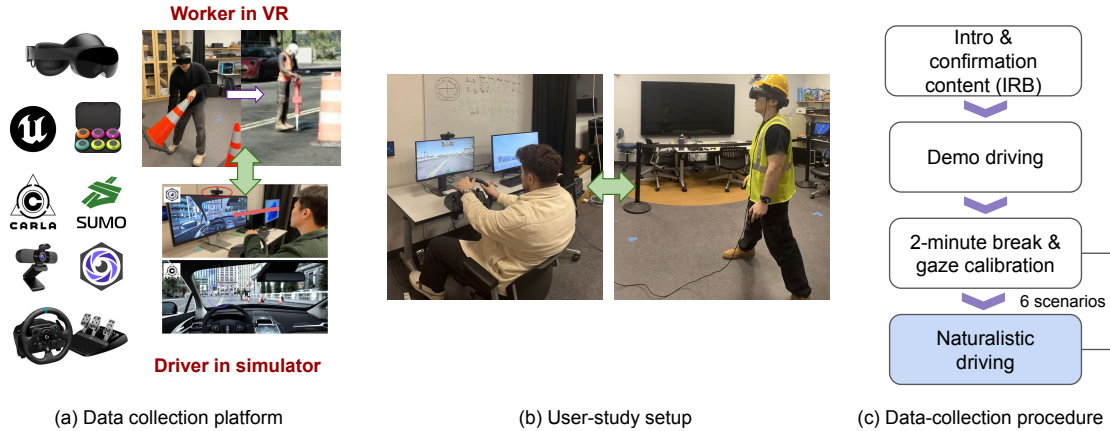


Figure 1: Overview of the DUALSENSE platform used to collect WZDRIVE-DATA. (a) Platform architecture connecting the human driver, VR roadway worker, simulation environment, and sensing devices. (b) User-study setup enabling real-time driver-worker interaction. (c) Data-collection procedure across six randomized scenario conditions.

six scenario conditions: a standard lane closure (S1, control), absent warning signage (S2), rain (S3), a curved road (S4), nighttime driving (S5), and truck involvement (S6). S2–S6 examine absent advance-warning signage, rain, curved roads, nighttime driving, and truck involvement. The weather, road geometry, lighting, and heavy-vehicle conditions reflect factors examined in empirical analyses of work-zone crashes [40–42]. The baseline layout comprised an 8 m coned work area with a warning sign 8 m upstream of the first cone; S2 omitted the sign. Each scenario used a distinct map with four to six SUMO-controlled vehicles, all trucks in S6. Their motion responded to surrounding traffic, including the ego vehicle, through the Intelligent Driver Model (IDM).

Participants and study procedure. We recruited 34 licensed drivers and excluded one participant’s data due to sensor disconnection. The final sample comprised 33 drivers aged 18–34 years ($M = 25.94$, $SD = 3.42$), with a median driving experience of 5.0 years ($Q_1 = 3.0$, $Q_3 = 7.25$). The study was approved by our institutional review board (IRB), whose identifying details are withheld for anonymous review. All participants provided informed consent and received compensation. Participants used a Logitech G923 steering wheel and pedals and were instructed to drive as they normally would. The study procedure is summarized in Fig. 1(c). Before recording, participants received instructions and completed gaze calibration and a familiarization drive. Experimental maps were previously unseen, and work-zone locations were not disclosed. The six scenarios were presented in randomized order, each lasting approximately 5 minutes and containing two encounters with each of WZ1–WZ3. A 2-minute break separated scenarios to reduce fatigue and carryover effects. After the experiment, participants completed a demographic questionnaire. Recordings were aligned by timestamp, and work-zone encounter segments were extracted to construct the dataset.

3.2 Dataset Characteristics

Dataset scale and temporal coverage. WZDRIVE-DATA comprises 1,183 distinct work-zone encounter segments from 33 drivers,

covering three work-zone configurations and six scenario conditions. Each segment corresponds to one participant, scenario, work zone, and lap. Together, these segments provide approximately 3.7 hours of trajectory coverage and yield 26,892 overlapping prediction windows. As summarized in Table 1, each window contains 10 observed frames and 25 future positions at 5 Hz, corresponding to a nominal 2 s observation history and a 5 s prediction horizon. Consecutive windows are separated by 0.2 s. Figure 2 shows their distribution across work zones and scenarios. Participant partitions are defined separately for each experimental protocol.

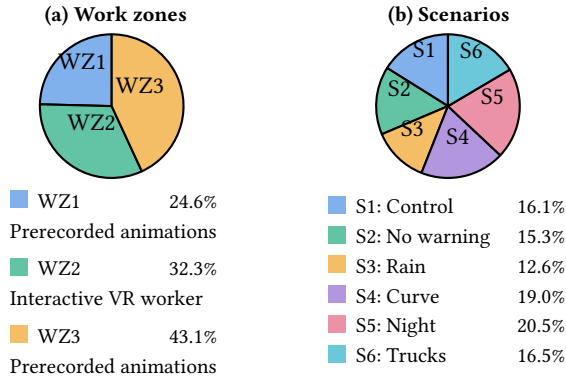
Multimodal content and representation. The structured inputs comprise six groups: ego motion, driver controls, gaze, surrounding agents, lane geometry, and work-zone context. Stored features contain 6 values per ego frame, 3 per control or gaze frame, and 11 per agent per frame, with up to 20 agents. Lane geometry is represented by 8-D point features, 10-D lane attributes, and graph connectivity. Work-zone reference points and static worker anchors use 3-D features encoding relative position and validity. Masks distinguish valid observations from missing or padded entries. In addition to these structured records, the dataset retains front-view RGB images, semantic-segmentation images, and driving-view screenshots. Gaze and all three visual streams are recorded at 10 Hz; gaze observations are aligned with the 5 Hz trajectory sequence when constructing model inputs.

Relation to general-driving datasets. For context, nuScenes contains 1,000 real-world scenes of 20 s each, totaling approximately 5.6 hours [6]. WZDRIVE-DATA adopts a nuScenes-style scene and agent organization, with trajectories represented as timestamped positions on the two-dimensional ground plane. However, nuScenes scenes and our work-zone encounter segments represent different collection units and should not be compared directly by count. Complementing broad real-world driving coverage, WZDRIVE-DATA focuses on repeatable work-zone conditions with synchronized driver controls, gaze, trajectories, and explicit temporary scene geometry.

Table 1: WZDRIVE-DATA scale and multimodality

(a) Dataset scale	
Characteristic	Value
Valid drivers	33
Work-zone configurations	3
Scenario conditions	6
Distinct encounter segments	1,183
Overlapping prediction windows	26,892
Trajectory coverage	≈ 3.7 h
Trajectory sampling rate	5 Hz
Observation / prediction horizon	2 s / 5 s
(b) Multimodal content	
Modality / scene information	Rate / representation
Ego-vehicle motion	5 Hz
Driver controls	5 Hz
Driver gaze	10 Hz
Surrounding-agent trajectories	5 Hz
Interactive VR-worker pose	5 Hz
Front-view RGB images	10 Hz
Semantic-segmentation images	10 Hz
Driving-view screenshots	10 Hz
Lane geometry and connectivity	Vector map
Work-zone geometry and worker animations	Scene context

Trajectory coverage counts overlapping windows only once within each encounter segment. Gaze is recorded at 10 Hz and aligned to 5 Hz for model input. Static geometry is represented spatially rather than assigned a sampling rate.

**Figure 2: Distribution of prediction windows by (a) work-zone configuration and (b) scenario condition.**

4 Methodology

We propose WZDRIVE, a multimodal ego-trajectory prediction model that combines driver behavior with roadway-work-zone context. The model encodes vehicle motion, driver-control history, gaze, and scene geometry to generate six candidate future trajectories and their probabilities. A temporal anchor-calibration module subsequently adjusts the predicted positions to address residual prediction errors. Figure 3 summarizes the encoder pretraining framework and subsequent training stages.

4.1 Problem Formulation

Given 2 s of observations, we predict the human-driven ego vehicle’s trajectory over the next 5 s. At 5 Hz, these windows contain $T_{\text{obs}} = 10$ observed frames and $T_{\text{pred}} = 25$ future frames. The input is

$$\mathcal{X} = \{\mathbf{X}^{\text{ego}}, \mathbf{X}^{\text{ctrl}}, \mathbf{X}^{\text{gaze}}, \mathbf{X}^{\text{agents}}, \mathcal{L}, \mathcal{W}\}, \quad (1)$$

where $\mathbf{X}^{\text{ego}}$ denotes ego-motion history; $\mathbf{X}^{\text{ctrl}}$ contains historical steering, throttle, and brake signals; and $\mathbf{X}^{\text{gaze}}$ denotes driver-gaze history. $\mathbf{X}^{\text{agents}}$ contains the motion histories of surrounding vehicles and VR-controlled workers. $\mathcal{L}$ represents lane geometry, while $\mathcal{W}$ contains work-zone geometry and static worker context. All temporal inputs are restricted to the observation window. Trajectory coordinates are expressed in a local frame centered at the ego vehicle’s latest observed position and aligned with its heading. To represent alternative future motions, WZDRIVE outputs $K = 6$ candidate trajectories:

$$\hat{\mathcal{Y}} = \{\hat{\mathbf{Y}}^{(1)}, \dots, \hat{\mathbf{Y}}^{(K)}\}, \quad \hat{\mathbf{Y}}^{(k)} \in \mathbb{R}^{T_{\text{pred}} \times 2}. \quad (2)$$

Each candidate contains 25 two-dimensional positions at 0.2 s intervals following the latest observation.

4.2 Multimodal Encoding and Fusion

Separate encoders process ego-motion history, driver-control history, gaze history, and surrounding-agent histories, producing $\mathbf{z}_{\text{motion}}$, $\mathbf{z}_{\text{control}}$, $\mathbf{z}_{\text{gaze}}$, and $\mathbf{z}_{\text{agents}}$, respectively. Lane geometry and work-zone context are encoded as $\mathbf{z}_{\text{lane}}$ and $\mathbf{z}_{\text{WZ}}$. The agent branch includes surrounding vehicles and VR-controlled workers, while the work-zone branch includes temporary geometry and static worker information.

Motion and control features are combined through learned scoring to form $\mathbf{z}_{\text{motion+control}}$. The scores weight control features to incorporate relevant driving actions into the motion representation, excluding invalid control observations. The model then uses $\mathbf{z}_{\text{motion+control}}$ to compute a relevance score for each agent and aggregate their features accordingly. Individual agent features are also retained so that the decoder can attend to specific vehicles and workers. Gaze is encoded separately and participates in the subsequent fusion.

Lane and work-zone representations are combined through a temporary topology adapter. Using lane features and $\mathbf{z}_{\text{WZ}}$, the adapter learns two types of scores: one assesses the suitability of each lane for travel under current work-zone conditions, while the other assesses the suitability of connections between lanes. Connection scores additionally consider the relative position, orientation, and connection type of the two lanes. These scores act as soft weights that jointly regulate information exchange within the lane graph. The updated lane features are pooled into $\mathbf{z}_{\text{map}}$, representing the road layout as affected by the work zone. Finally, horizon-aware fusion combines the four representations:

$$\mathbf{c}_h = \mathcal{F}_h(\mathbf{z}_{\text{motion+control}}, \mathbf{z}_{\text{gaze}}, \mathbf{z}_{\text{agents}}, \mathbf{z}_{\text{map}}). \quad (3)$$

Here, $h \in \{1, 3, 5\}$ denotes the prediction horizon in seconds. The contexts $\mathbf{c}_1$, $\mathbf{c}_3$, and $\mathbf{c}_5$ guide predictions over future intervals $(0, 1]$, $(1, 3]$, and $(3, 5]$ s, respectively. Each context uses horizon-dependent modality weights, allowing different intervals to draw on different information combinations. The decoder predicts the complete 5 s

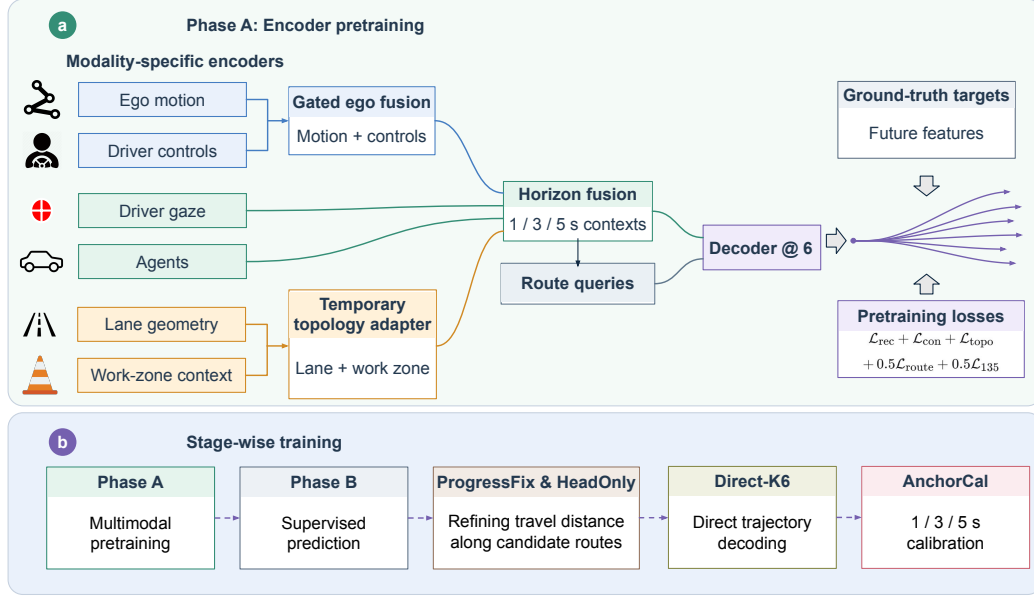


Figure 3: Overview of WZDRIVE pretraining and staged training. (a) Modality-specific encoders and fusion modules learn representations through reconstruction, contrastive, topology, route-coverage, and travel-distance objectives. (b) The six-stage training schedule, from encoder pretraining to temporal anchor calibration.

trajectory at 0.2 s intervals, yielding 25 positions rather than only three endpoints.

Trajectory decoder. Inspired by MTR’s use of mode-specific queries [22], the decoder generates six candidate trajectories over 5 s at 5 Hz, each containing 25 ego-local positions at 0.2 s intervals. A learned motion-prior network uses $\mathbf{z}_{\text{motion+control}}$ and $\mathbf{z}_{\text{control}}$ to estimate all 25 future velocities. Multiplying these by $\Delta t = 0.2$ s yields prior displacements $\Delta \mathbf{y}_t^{\text{prior}}$ shared across candidates.

Each candidate has 25 time-indexed queries combining motion-control features, a candidate embedding, time encoding, the corresponding fused context $\mathbf{c}_h$, and encoded prior position and velocity. Cross-attention updates these queries using ego, lane, agent, worker, and work-zone features, with invalid features masked. All candidates and future steps are processed in a single forward pass, rather than autoregressively, to predict candidate-specific displacement residuals $\mathbf{r}_{k,t}$. Adding these residuals to the shared prior displacements and accumulating in temporal order gives the initial predicted positions:

$$\hat{\mathbf{y}}_{k,t}^{\text{init}} = \sum_{\tau=1}^t (\Delta \mathbf{y}_{\tau}^{\text{prior}} + \mathbf{r}_{k,\tau}). \quad (4)$$

Here, $k \in \{1, \dots, 6\}$ indexes candidates and $t \in \{1, \dots, 25\}$ indexes future steps. A scoring head assigns candidate probabilities, and temporal anchor calibration further adjusts the initial trajectories (Section 4.3).

4.3 Stage-wise Training and Trajectory Refinement

Training proceeds from multimodal pretraining to supervised prediction and subsequent refinement. Early stages use candidate

routes through the lane graph, with route progress specifying how far the vehicle travels along each route. Later stages introduce the direct trajectory decoder and calibrate its predictions.

Phase A: Multimodal pretraining. Inspired by masked pretraining [25] and contrastive learning for trajectory prediction [43], we pretrain the encoders, fusion modules, and route-prediction components using

$$\mathcal{L}_{\text{PhaseA}} = \mathcal{L}_{\text{rec}} + \mathcal{L}_{\text{con}} + \mathcal{L}_{\text{topo}} + 0.5\mathcal{L}_{\text{route}} + 0.5\mathcal{L}_{135}. \quad (5)$$

The reconstruction loss $\mathcal{L}_{\text{rec}}$ teaches the model to recover masked input information. The contrastive loss $\mathcal{L}_{\text{con}}$ helps the model associate observed driving context with its corresponding future behavior. The topology loss $\mathcal{L}_{\text{topo}}$ supervises lane connections and the spatial relationships between lanes and work-zone elements. The route-coverage loss $\mathcal{L}_{\text{route}}$ encourages candidate routes to include a path close to the driver’s actual route. The travel-distance loss $\mathcal{L}_{135}$ supervises the cumulative distance the vehicle travels by 1, 3, and 5 s.

Phase B: Supervised trajectory prediction. Whereas Phase A learns representations and route information, Phase B trains complete trajectory predictions through direct position supervision:

$$\mathcal{L}_{\text{PhaseB}} = \mathcal{L}_{\text{traj}} + 0.5\mathcal{L}_{\text{end}} + \mathcal{L}_{\text{aux}} + 0.2\mathcal{L}_{\text{local-refine}}. \quad (6)$$

The trajectory loss $\mathcal{L}_{\text{traj}}$ penalizes position errors at all 25 future steps, while the endpoint loss $\mathcal{L}_{\text{end}}$ emphasizes the final position. Both additionally penalize longitudinal errors. The auxiliary loss $\mathcal{L}_{\text{aux}}$ combines weighted objectives for route prediction, candidate ranking, and consistency with lanes, work zones, and surrounding agents. The refinement loss $\mathcal{L}_{\text{local-refine}}$ trains the refiner to correct coarse trajectory position errors. The trajectory and endpoint losses are reused in subsequent training stages.

Table 2: Rule-based and time-series baseline performance ($K = 1$) with selected input combinations. Lower is better.

Model	Features	ADE (m) ↓	FDE (m) ↓
Stay	Ego	21.635	42.053
CV	Ego	7.237	16.896
LSTM	Ego + control + WZ	2.988	7.513
Transformer	Ego + control + WZ	2.876	7.284
GRU	Ego + control + WZ + lane	2.860	7.320

Route-progress refinement. Inspired by PGP’s route-conditioned prediction [18], we refine route progress in two stages. **ProgressFix** refines the predicted distance traveled along a candidate route by 1, 3, and 5 s using $\mathcal{L}_{\text{progress}}$. **HeadOnly** further refines these predictions by learning corrections to the travel increment at each of the 25 future steps.

Direct-K6: Direct trajectory prediction. Direct-K6 refines six-candidate trajectory prediction to improve accuracy and promote diversity among candidate routes. Starting from the HeadOnly checkpoint, we introduce a new direct decoder and jointly train it with the existing modules. The training objective combines trajectory loss $\mathcal{L}_{\text{traj}}$, endpoint loss $\mathcal{L}_{\text{end}}$, candidate-scoring loss $\mathcal{L}_{\text{score}}$, motion-prior loss $\mathcal{L}_{\text{prior}}$, and route-diversity loss $\mathcal{L}_{\text{div}}$.

AnchorCal: Temporal position calibration. Similar to the residual trajectory refinement in LAFormer [26], AnchorCal learns position corrections to improve the initial predictions. We freeze the Direct-K6 predictor and attach a calibration head to its decoder features. The head predicts longitudinal and lateral position corrections at 1, 3, and 5 s. These corrections are linearly interpolated to all 25 future steps and added to the original predictions:

$$\hat{\mathbf{y}}_{k,t}^{\text{final}} = \hat{\mathbf{y}}_{k,t}^{\text{init}} + \delta_{k,t}. \quad (7)$$

Here, $\delta_{k,t}$ denotes the interpolated position correction, with zero correction at the current time. The calibrated trajectories are compared with the ground truth using $\mathcal{L}_{\text{traj}}$ and $\mathcal{L}_{\text{end}}$, updating only the calibration head.

5 Results

5.1 Experimental Setup

Protocol and implementation. All models observe 2 s and predict 5 s of ego motion at 5 Hz. The main comparison uses participant-disjoint Set1: 27 training drivers, P14/P20/P34 for validation, and P2/P12/P31 for testing, yielding 22,540/2,119/2,233 windows, respectively. Experiments used an Intel Core i7-14700K CPU, 64 GB RAM, and an NVIDIA RTX 4090 GPU.

Baselines and metrics. We compare rule-based Stay/CV, GRU, LSTM, Transformer, and our MM-GRU baseline. Advanced methods represent visual and gaze modeling (RouteFormer), lane-aware prediction (LAFormer, PGP), query-based prediction (MTR), intention-dynamics decomposition (DeMo), and masked pretraining (ForecastMAE); citations are provided in Table 3. We evaluate WZ and no-WZ variants where available. Single-trajectory baselines use ADE/FDE; six-candidate models use minADE₆/minFDE₆, which independently select the lowest average trajectory and endpoint errors. Results are averaged over test samples in meters; lower is better.

5.2 Trajectory Prediction Performance

Rule-based and time-series baselines. Table 2 compares single-trajectory predictors ($K = 1$), including stationary (Stay), constant-velocity (CV), and time-series models. For each learned model, we report its best-performing input combination among those evaluated. All learned models outperform Stay and CV. GRU achieves the lowest ADE, while Transformer achieves the lowest FDE, reflecting both architecture and input differences.

Multimodal forecasting baselines. Table 3 compares methods standardized to six candidate trajectories ($K = 6$). We evaluate variants with and without work-zone information, incorporating it through an additional encoder or the existing lane encoder and reporting the best-performing integration evaluated for each model. Work-zone information reduces minADE for every paired baseline and minFDE for all except DeMo, demonstrating its value for prediction in this setting.

Among published baselines, LAFormer achieves the lowest minADE, consistent with its lane-aware design, while DeMo achieves the lowest minFDE. RouteFormer has the highest errors on both metrics. Its reliance on RGB imagery and gaze may make learning more demanding under our limited-data setting. Motivated by GRU’s competitive time-series performance, we develop MM-GRU to predict six candidate trajectories and evaluate different input combinations. The selected configuration combines ego motion, driver controls, surrounding agents, lane geometry, and work-zone context, including worker information. MM-GRU outperforms the published baselines, supporting the utility of multimodal inputs. Together, these results suggest that effective representation and fusion matter more than simply adding modalities. MM-GRU outperforms the published baselines, supporting the utility of multimodal inputs. Together, these results suggest that effective representation and fusion matter more than simply adding modalities.

WZDrive achieves the lowest mean errors across three seeds: 1.094 ± 0.022 m minADE₆ and 2.115 ± 0.082 m minFDE₆. Relative to the best published baseline for each metric, it reduces minADE by 7.18% compared with LAFormer with WZ and minFDE by 13.18% compared with DeMo without WZ.

5.3 Ablation Studies

Input-modality ablation. We mask the selected inputs and retrain each variant from scratch using the same training schedule. The WZ variant also disables the associated topology adaptation and auxiliary losses. Table 4(a) shows that removing gaze has little effect in this setting. Removing driver controls causes the largest degradation, increasing minADE and minFDE by 17.20% and 14.17%, respectively. Removing WZ and worker context and its associated mechanisms increases these errors by 7.54% and 12.78%, supporting the value of human control signals and work-zone-aware modeling. **Stage-wise training.** Table 4(b) shows the effectiveness of the overall training pipeline. Although intermediate stages do not improve monotonically, Direct-K6 produces the largest error reduction, and AnchorCal further improves most metrics. The final model outperforms Phase B across all three participant partitions.

Participant generalization. We additionally evaluate two participant-disjoint partitions: Set2 uses P5/P15/P18 for validation and P3/P4/P25 for testing; Set3 uses P7/P11/P19 and P13/P17/P23, respectively. The

Table 3: Comparison with SOTA baselines ($K = 6$). Errors are in meters.

Model	Without WZ		With WZ	
	minADE ₆ ↓	minFDE ₆ ↓	minADE ₆ ↓	minFDE ₆ ↓
PGP (2021) [18]	1.286	2.749	1.260 (↓2.02%)	2.742 (↓0.25%)
MTR (2022) [22]	1.317	2.660	1.299 (↓1.37%)	2.567 (↓3.50%)
Forecast-MAE (2023) [25]	1.373	2.650	1.264 (↓7.94%)	2.475 (↓6.60%)
LAFormer (2024) [26]	<u>1.204</u>	2.662	<u>1.179</u> (↓2.08%)	2.550 (↓4.21%)
DeMo (2024) [27]	1.514	<u>2.436</u>	1.281 (↓15.39%)	<u>2.462</u> (↑1.07%)
RouteFormer (2025) [28]	2.106	4.023	2.050 (↓2.65%)	3.847 (↓4.38%)
MM-GRU (ours)	1.234	2.682	1.124 (↓8.91%)	2.242 (↓16.41%)
WZDRIVE (ours) [†]	–	–	1.094 ± 0.022 (↓7.18%)	2.115 ± 0.082 (↓13.18%)

Baseline percentages compare with the same model without WZ. [†]WZDrive percentages compare with the best published baselines: LAFormer with WZ for minADE₆ and DeMo without WZ for minFDE₆. WZDrive reports mean ± SD over three seeds. Underlining marks the best published baseline per column, excluding MM-GRU; bold marks the best overall means.

Table 4: Effects of input modalities and training stages across participant partitions. Errors are in meters; lower is better.

(a) Input modalities on Set1			
Variant	minADE ₆	minFDE ₆	
WZDRIVE (full)	1.103	2.119	
Without gaze	1.108	2.116	
Without driver controls	1.293	2.419	
Without WZ and workers	1.187	2.390	

(b) Training stages across participant partitions			
Stage	Set1	Set2	Set3
Phase B	1.572 / 3.692	1.900 / 4.156	1.989 / 4.133
ProgressFix	1.621 / 3.827	1.910 / 4.207	1.981 / 4.377
HeadOnly	1.591 / 3.614	1.913 / 4.169	2.001 / 4.341
Direct-K6	1.148 / 2.244	1.418 / 2.749	1.530 / 2.801
AnchorCal	1.094 / 2.115	1.345 / 2.539	1.496 / 2.809

(a) Each input variant is retrained from scratch. (b) Entries report minADE₆ / minFDE₆ across three participant partitions. AnchorCal is the final WZ-DRIVE model.

remaining 27 participants form each training partition. Both show an overall error-reduction trend similar to Set1, while differences in absolute performance may reflect variation in driving behavior across participant groups.

6 Discussion

Gaze and driver behavior. Although gaze provides limited additional improvement in our ablation, this does not mean that driver attention is unimportant. The five safety-critical scenarios involve challenges arising from scenario design, temporary geometry, and VR-worker interactions, rather than ordinary driving conditions alone. In these settings, visual attention is relevant to how drivers

perceive their surroundings, make decisions, and adjust their driving behavior. Previous work-zone studies have similarly examined the relationship between gaze, driver awareness, and worker behavior [38, 39]. We also explored a model combining semantic images with gaze instead of RGB images, but its performance remained below WZDRIVE. Together, these observations suggest that the behavioral value of gaze is not automatically translated into prediction gains: effectively incorporating attention and visual context remains important for multimodal trajectory prediction.

Stage-wise training and trajectory refinement. The stage-wise results in Table 4(b) highlight the value of the training and trajectory refinement described in Section 4.3. Route-progress refinement adjusts travel along candidate routes, while Direct-K6 introduces direct prediction of six trajectories using a shared motion prior. AnchorCal then adds learned position corrections at 1, 3, and 5 s, interpolated across all future steps. Although improvements are not monotonic, Direct-K6 produces the largest error reduction. Relative to Direct-K6, AnchorCal further reduces minADE₆ across all three partitions and minFDE₆ in Set1 and Set2, while minFDE₆ in Set3 increases slightly. Similar overall trends across participant partitions support the effectiveness of the training pipeline.

Representative failure cases. Figure 4 shows five cases selected from ten high-error candidates across the test partitions of Set1–Set3. Each represents a distinct encounter, excluding repeated sliding windows from the same trajectory. The top row shows driving views with gaze overlays; green markers indicate high-confidence gaze measurements. The bottom row shows observed trajectories, ground-truth futures, and predictions.

Based on video inspection, the first case shows a collision near the work zone with an unclear cause. In the second, the driver attempts to overtake a SUMO-controlled vehicle blocking the view of the work zone. Missing advance-warning signage in S2 further limits early warning before a work-zone collision. In the third, the driver stops to observe the work zone without a collision or interference from another vehicle, possibly reflecting caution in

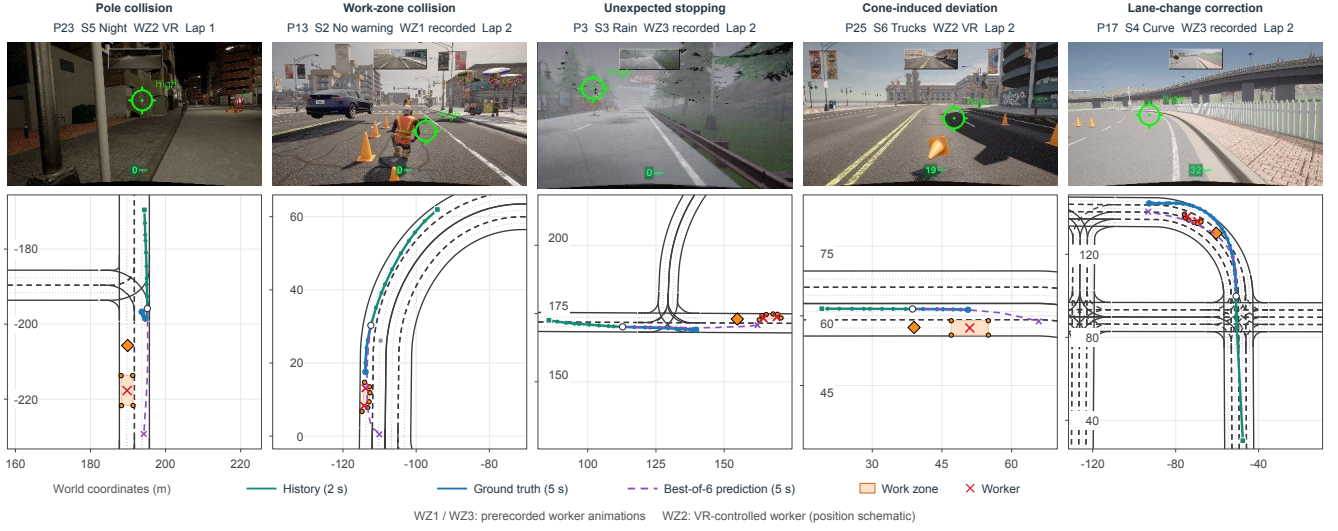


Figure 4: Qualitative failure cases. From left to right: pole collision, work-zone collision, unexpected stopping, cone-induced deviation, and lane-change correction. The top row shows driving views with gaze overlays; the bottom row shows observed motion, ground-truth futures, and the lowest-ADE prediction among six candidates. VR-worker markers indicate schematic positions.

rain. In the fourth, a cone enters the roadway during a Lap 2 VR-worker interaction, triggering abrupt braking. Some hazardous interactions were intentionally introduced in Lap 2 to enrich the dataset. Finally, in the fifth case, the curve and nearby work zone make a lane-change maneuver difficult, causing the vehicle to drive onto the curb.

These cases highlight the difficulty of predicting abrupt or unusual driving responses near work zones. Although the scenarios were designed, the specific collisions, stops, and corrective actions arose from human driving rather than being prescribed. These observations motivate broader training coverage of such responses and highlight the value of simulation for collecting them safely.

Applications and pathway to deployment. WZDrive could support ADAS in roadway work zones by informing warnings and corrective interventions when drivers make control errors. Its predictions could also inform downstream planning and control modules in autonomous driving. For Level 2–3 applications, the control history records steering, throttle, and braking inputs over the preceding 2 s, originating from either the driver or the driving system depending on who controls the vehicle. During Level 4–5 autonomous operation, these inputs are generated by the vehicle controller rather than a human driver. For on-vehicle deployment, historical control signals can be obtained through the Controller Area Network (CAN) bus and decoded into numerical features. Vehicle-specific pre-processing aligns their definitions, units, ranges, and sampling rates with the model inputs. The gaze would be supplied separately by an eye-tracking system and could be optional depending on the driver.

7 Conclusion and Future Work

This study addresses the prediction of the ego-trajectory in urban roadway work zones. Unlike regular driving environments, temporary road work and associated worker activities change how vehicles navigate and interact with their surroundings. Studying prediction under these conditions requires data that capture both work-zone context and human driving responses. We introduce WZDRIVE-DATA, collected through human-in-the-loop driving and VR-worker interaction, establish a trajectory-prediction benchmark with adapted baselines, and develop WZDRIVE, a multimodal predictor with stage-wise training and refinement. The results demonstrate the value of driver-control signals and explicit work-zone context for this task.

Future work will address two main directions. First, the current data were collected in a simulation tool developed by the authors as this paper from 33 drivers aged 18–34. Broader recruitment across age groups and driving experience, together with real-world data, is needed to assess generalization. Second, although the data set covers three work-zone configurations and six scenarios, rare events such as collisions, atypical stopping, and unexpected worker interactions remain sparsely represented. Expanding work-zone layouts and interaction conditions could provide additional examples of these events for more comprehensive training and evaluation.

Overall, the simulated data set (WZDRIVE-DATA), the work-zone ego-trajectory prediction benchmark, and the model (WZDRIVE) provide a foundation for studying ego-vehicle trajectory prediction in urban work zones. The human-interaction-based data collection approach could also be extended to settings such as vehicle-pedestrian encounters, supporting the study of driving responses in complex environments and the development of autonomous vehicles that can better navigate work zones.

References

- [1] Work Zone Safety Information Clearinghouse. Work zone data, 2026. Data derived from NHTSA FARS and CRSS; accessed September 3, 2026.
- [2] National Highway Traffic Safety Administration. Crash report sampling system (CRSS). Accessed September 3, 2026.
- [3] Gerald L. Ullman, Jeremy Schroeder, and Deepak Gopalakrishna. Use of technology and data for effective work zone management: Work zone ITS implementation guide. Technical Report FHWA-HOP-14-008, Federal Highway Administration, January 2014.
- [4] Isht Dwivedi, Srikanth Malla, Yi-Ting Chen, and Behzad Dariush. Bird's eye view segmentation using lifted 2d semantic features. In *Proceedings of the British Machine Vision Conference*, 2021.
- [5] Anurag Ghosh, Shen Zheng, Robert Tamburo, Khiem Vuong, Juan R. Alvarez-Padilla, Hailiang Zhu, Michael Cardei, Nicholas Dunn, Christoph Mertz, and Srinivasa G. Narasimhan. ROADWork: A dataset and benchmark for learning to recognize, observe, analyze and drive through work zones. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6132–6142, 2025.
- [6] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuScenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11621–11631, 2020.
- [7] Ming-Fang Chang, John Lambert, Patsorn Sangkloy, Jagjeet Singh, Slawomir Bak, Andrew Hartnett, De Wang, Peter Carr, Simon Lucey, Deva Ramanan, and James Hays. Argoverse: 3d tracking and forecasting with rich maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8748–8757, 2019.
- [8] Scott Etinger, Shuyang Cheng, Benjamin Caine, Chenxi Liu, Hang Zhao, Sabeek Pradhan, Yuning Chai, Ben Sapp, Charles R. Qi, Yin Zhou, Zoey Yang, Aurick Zhou, Pei Sun, Jiquan Ngiam, Vijay Vasudevan, Alexander McCauley, Jonathon Shlens, and Dragomir Anguelov. Large scale interactive motion forecasting for autonomous driving: The waymo open motion dataset. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9710–9719, 2021.
- [9] Gustavo Silveira, Abhijat Biswas, and Henny Admoni. DReyeVR: Democratizing virtual reality driving simulation for behavioural & interaction research. In *Proceedings of the 2022 ACM/IEEE International Conference on Human-Robot Interaction*, pages 639–643, 2022.
- [10] Zhuoli Zhuang, Yu-Cheng Chang, Yu-Kai Wang, Thomas Do, and Chin-Teng Lin. Neuro-cognitive reward modeling for human-centered autonomous vehicle control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10599–10609, June 2026.
- [11] Zhuoli Zhuang, Cheng-You Lu, Yu-Cheng Fred Chang, Yu-Kai Wang, Thomas Do, and Chin-Teng Lin. AEGIS: Human attention-based explainable guidance for intelligent vehicle systems, 2025.
- [12] Abhijat Biswas, Badal Arun Pardihi, Caleb Chuck, Jarrett Holtz, Scott Niekum, Henny Admoni, and Alessandro Allievi. Gaze supervision for mitigating causal confusion in driving agents. In *2024 IEEE Intelligent Vehicles Symposium (IV)*. IEEE, 2024.
- [13] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 1–16, 2017.
- [14] Pablo Alvarez Lopez, Michael Behrisch, Laura Bieker-Walz, Jakob Erdmann, Yun-Pang Flötteröd, Robert Hilbrich, Leonhard Lücken, Johannes Rummel, Peter Wagner, and Evamarie Wießner. Microscopic traffic simulation using SUMO. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*, pages 2575–2582. IEEE, 2018.
- [15] Tim Salzmann, Boris Ivanovic, Punarjay Chakravarty, and Marco Pavone. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *European Conference on Computer Vision*, pages 683–700, 2020.
- [16] Abdullah Mohamed, Kun Qian, Mohamed Elhoseiny, and Christian Claudel. Social-STGCNN: A social spatio-temporal graph convolutional neural network for human trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14424–14432, 2020.
- [17] Ming Liang, Bin Yang, Rui Hu, Yun Chen, Renjie Liao, Song Feng, and Raquel Urtasun. Learning lane graph representations for motion forecasting. In *European Conference on Computer Vision*, pages 541–556, 2020.
- [18] Nachiket Deo, Eric Wolff, and Oscar Beijbom. Multimodal trajectory prediction conditioned on lane-graph traversals. In *Proceedings of the 5th Conference on Robot Learning*, volume 164 of *Proceedings of Machine Learning Research*, pages 203–212. PMLR, 2022.
- [19] Zikang Zhou, Luyao Ye, Jianping Wang, Kui Wu, and Kejie Lu. HiVT: Hierarchical vector transformer for multi-agent motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8823–8833, 2022.
- [20] Ye Yuan, Xinshuo Weng, Yanglan Ou, and Kris M. Kitani. AgentFormer: Agent-aware transformers for socio-temporal multi-agent forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9813–9823, 2021.
- [21] Nigamaa Nayakanti, Rami Al-Rfou, Aurick Zhou, Kratharth Goel, Khaled S. Refaat, and Benjamin Sapp. Wayformer: Motion forecasting via simple and efficient attention networks. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 2980–2987, 2023.
- [22] Shaoshuai Shi, Li Jiang, Dengxin Dai, and Bernt Schiele. Motion transformer with global intention localization and local movement refinement. In *Advances in Neural Information Processing Systems*, volume 35, 2022.
- [23] Shaoshuai Shi, Li Jiang, Dengxin Dai, and Bernt Schiele. MTR++: Multi-agent motion prediction with symmetric scene modeling and guided intention querying. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3955–3971, 2024.
- [24] Zikang Zhou, Jianping Wang, Yung-Hui Li, and Yu-Kai Huang. Query-centric trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17863–17873, 2023.
- [25] Jie Cheng, Xiaodong Mei, and Ming Liu. Forecast-MAE: Self-supervised pre-training for motion forecasting with masked autoencoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8679–8689, 2023.
- [26] Mengmeng Liu, Hao Cheng, Lin Chen, Hellward Broszio, Jiangtao Li, Runjiang Zhao, Monika Sester, and Michael Ying Yang. LAFormer: Trajectory prediction for autonomous driving with lane-aware scene constraints. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 2039–2049, 2024.
- [27] Bozhou Zhang, Nan Song, and Li Zhang. DeMo: Decoupling motion forecasting into directional intentions and dynamic states. In *Advances in Neural Information Processing Systems*, 2024.
- [28] M. Eren Akbiyik, Nedko Savov, Danda Pani Paudel, Nikola Popovic, Christian Vater, Otmar Hilliges, Luc Van Gool, and Xi Wang. Leveraging driver field-of-view for multimodal ego-trajectory prediction. In *International Conference on Learning Representations*, 2025.
- [29] Stefano Alletto, Andrea Palazzi, Francesco Solera, Simone Calderara, and Rita Cucchiara. DR(Eye)VE: A dataset for attention-based tasks with applications to autonomous and assisted driving. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 54–60, 2016.
- [30] Young-Woo Seo, JongHo Lee, Wende Zhang, and David Wettergreen. Recognition of highway workzones for reliable autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 16(2):708–718, 2015.
- [31] Anurag Ghosh, Francesco Pittaluga, Khiem Vuong, Angela Chen, Juan R. Alvarez-Padilla, Manmohan Chandraker, and Srinivasa G. Narasimhan. Dash2Sim: Closed-loop driving simulation from in-the-wild dashcam videos. *arXiv preprint arXiv:2606.07366*, 2026.
- [32] Erica Weng, Kenta Mukoya, Deva Ramanan, and Kris Kitani. Evaluating a VR system for collecting safety-critical vehicle–pedestrian interactions. *arXiv preprint arXiv:2310.05882*, 2023.
- [33] Kenta Mukoya, Erica Weng, Rohan Choudhury, and Kris Kitani. JaywalkerVR: A VR system for collecting safety-critical pedestrian–vehicle interactions. *arXiv preprint arXiv:2407.04843*, 2024.
- [34] Luca Crosato, Chongfeng Wei, Edmond S. L. Ho, Hubert P. H. Shum, and Yuzhu Sun. A virtual reality framework for human-driver interaction research: Safe and cost-effective data collection. In *Proceedings of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pages 167–174, 2024.
- [35] Meiting Dang, Yan Jin, Peng Hang, Luca Crosato, Yuzhu Sun, and Chongfeng Wei. Coupling intention and actions of vehicle–pedestrian interaction: A virtual reality experiment study. *Accident Analysis & Prevention*, 203:107639, 2024.
- [36] Meiting Dang, Dezong Zhao, Yafei Wang, and Chongfeng Wei. Dynamic game-theoretical decision-making framework for vehicle–pedestrian interaction with human bounded rationality. *IEEE Transactions on Intelligent Transportation Systems*, 26(7):10822–10833, 2025.
- [37] Zhaobo Zheng, Kumar Akash, and Teruhisa Misu. DualTake: Predicting takeovers across mobilities for future personalized mobility services. In *Companion of the 2024 ACM/IEEE International Conference on Human-Robot Interaction*, pages 1194–1198, 2024.
- [38] Shuo Zhang, Fan Zuo, Semiha Ergun, and Kaan Ozbay. Improving work zone safety: Integrating VR-CARLA co-simulation and eye tracking for behavior analysis of drivers around work zones. *Journal of Transportation Safety & Security*, 2025.
- [39] Shuo Zhang, Zu Wang, Semiha Ergun, and Kaan Ozbay. VR-WISE: VR-based work zone immersive simulation with eye tracking and biometric sensors for capturing driver's awareness. In *Proceedings of the 17th International Conference on Automotive User Interfaces and Interactive Vehicular Applications*, pages 69–79, 2025.
- [40] Ali Ghasemzadeh and Mohamed M. Ahmed. Exploring factors contributing to injury severity at work zones considering adverse weather conditions. *IATSS Research*, 43(3):131–138, 2019.
- [41] Yingfeng Li and Yong Bai. Highway work zone risk factors and their impact on crash severity. *Journal of Transportation Engineering*, 135(10):694–701, 2009.
- [42] Mouyid Islam. Assessing the effects of geometry and non-geometry related factors in work-zone crashes. *Traffic Injury Prevention*, 25(3):492–498, 2024.

- [43] Pengfei Yao, Yinglong Zhu, Huikun Bi, Tianlu Mao, and Zhaoqi Wang. TrajCLIP: Pedestrian trajectory prediction method using contrastive learning and idempotent networks. In *Advances in Neural Information Processing Systems*, volume 37, 2024.